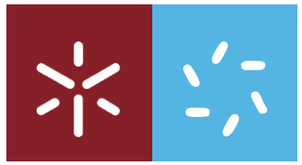
Modelos Longitudinais e de Sobrevivência para a Recidiva do Cancro da Mama

Liliana Rodrigues Coelho

UMinho | 2017



**Universidade do Minho**
Escola de Ciências

Liliana Rodrigues Coelho

# Modelos Longitudinais e de Sobrevivência para a Recidiva do Cancro da Mama

Janeiro de 2017

Universidade do Minho
Escola de Ciências

Liliana Rodrigues Coelho

# Modelos Longitudinais e de Sobrevivência para a Recidiva do Cancro da Mama

Tese de Mestrado

Estatística

Trabalho efetuado sob a orientação da

**Professora Doutora Inês Sousa**

e coorientação da

**Professora Doutora Ana Borges**

Janeiro de 2017

# Agradecimentos

Não quero terminar esta tese sem agradecer às pessoas que neste ano me acompanharam, quer física quer psicologicamente.

Quero agradecer, em primeiro lugar, às duas orientadoras, a Professora Doutora Inês Sousa, e a Professora Doutora Ana Borges, por quem tenho um grande carinho e admiração, pela disponibilidade e simpatia demonstradas em todos os momentos, assim como pela partilha de conhecimentos e pela exigência no trabalho efetuado.

Às minhas duas colegas de trabalho, Ana Paula Sousa e Ana Teresa Torre, pela ajuda que me deram na correção e eliminação dos meus erros, quer na língua inglesa, quer na língua portuguesa, respetivamente.

Aos meus colegas, em especial à Beatriz, pelo apoio e pela amizade.

Aos meus professores, que me acompanharam no percurso académico; ao professor Luís Machado, pelo acompanhamento e esclarecimento das dúvidas pontuais; à professora Susana Faria, pela disponibilidade demonstrada para a realização bem-sucedida à sua disciplina; à professora Raquel Menezes, pela sua simpatia, e à professora Inês Sousa, pela sua paciência incondicional, pelo seu profissionalismo e por ter confiado em mim como sua orientada.

Por último, um profundo agradecimento à minha família, em especial à minha irmã Helena, que me apoiou incondicionalmente nesta minha aventura.

A todos, o meu sincero agradecimento.

# Resumo

O interesse por esta temática deve-se à complexidade e mistério que ainda envolve uma doença que nos afeta, principalmente a nós mulheres e que, infelizmente, não obstante a ciência ter evoluído imenso na área da investigação, continua ainda a intrigar-nos pela evolução e desfecho que a própria doença envolve.

O objetivo é inferir sobre os fatores de risco que afetam a sobrevivência para a recidiva do cancro da mama.

Analisando também os fatores de risco que afetam a progressão no tempo de dois marcadores tumorais, o marcador CA15.3 e o marcador CEA.

Para a análise de sobrevivência recorreu-se ao estimador de Kaplan-Meier e ao modelo de riscos proporcionais de Cox.

Os modelos lineares generalizados foram utilizados para estudar a evolução dos marcadores tumorais. Após análise individual das variáveis ajustamos os dois modelos longitudinais multivariados, considerando o efeito conjunto de covariáveis, um para cada marcador, adotando a metodologia de efeitos aleatórios

Concluímos que as variáveis, triplo negativo, tamanho do tumor e idade ao diagnóstico têm efeito na probabilidade de recidiva do tumor.

Na progressão do marcador CA15.3, as variáveis, idade ao diagnóstico, imagens de invasão vascular venosa e Ki.67 têm efeito na progressão, ao longo do tempo, desde a data do diagnóstico até à data do teste, enquanto, a variável imagem de invasão vascular tem efeito na progressão do marcador ao longo do tempo, desde a data da recidiva até à data do teste.

Relativamente à progressão do marcador CEA, ao longo do tempo, desde a data do diagnóstico até à data do teste, as variáveis idade ao diagnóstico e estadio apresentam um efeito significativo. Considerando o tempo, desde a data da recidiva até à data do teste, as variáveis bilateral e estadio, têm influência na progressão deste marcador.

**Palavras-Chave:** Cancro da mama, Análise de Sobrevivência, Análise Longitudinal.

# Abstract

The interest in this subject is due to the complexity and mystery that still surrounds a disease that affects us, mainly to us women and that, unfortunately, although the science has advanced immensely in the area of the investigation, still continues to intrigue us by the evolution and the outcome that the disease itself involves.

The objective is to infer about the risk factors that affect survival for recurrence of breast cancer.

Analyzing also the risk factors that affect the time progression of two tumor markers, the CA15.3 marker and the CEA marker.

For the survival analysis, we used the Kaplan-Meier estimator and the Cox proportional hazards model.

Generalized linear models were used to study the evolution of tumor markers. After individual analysis of the variables, we adjusted the two multivariate longitudinal models, considering the covariates set, one for each marker, adopting the random effects methodology

We conclude that the variables, triple negative, tumor size and age at diagnosis have an effect on the probability of tumor recurrence.

In the progression of the CA15.3 marker, variables, age at diagnosis, venous vascular invasion images, and Ki.67 have an effect on progression over time from the date of diagnosis to the date of the test, while the variable Image of vascular invasion has effect on marker progression over time, from the date of relapse to the test date.

Regarding the progression of the CEA marker, over time, from the date of diagnosis to the date of the test, the variables age at diagnosis and stage have a significant effect. Considering the time from the date of relapse to the date of the test, bilateral and stage variables have an influence on the progression of this marker.



**Key Words:** Breast Cancer, Survival Analysis, Longitudinal Analysis.

# ÍNDICE

# Índice de Tabelas

# Índice de figuras

# Capítulo 1- Introdução

Este trabalho tem como intuito desenvolver uma análise de dados de tempos de sobrevivência até à recidiva de pacientes que foram diagnosticados com cancro da mama, no período de 2008 até 2013, no Hospital de Braga, recorrendo a métodos de análise de sobrevivência e análise longitudinal. Há o interesse de entender as progressões dos marcadores tumorais até à data da recidiva, bem como o que acontece imediatamente antes da recidiva.

A Análise de Sobrevivência é a metodologia adequada para modelar um estudo quando se tem como variável resposta o tempo até à ocorrência de um evento de interesse, sendo este denominado tempo de falha.

Além deste tempo de falha, consideramos também o tempo de censura. O tempo de censura descreve o processo dos indivíduos que, por algum motivo, não concluíram o evento de interesse.

Neste trabalho o modelo de Cox será considerado, por ser um modelo semi-paramétrico, isto é, por ter uma componente paramétrica e uma componente não-paramétrica.

Este estudo está organizado da seguinte forma:

No capítulo 2, apresenta-se uma breve descrição do Hospital de Braga, assim como uma breve descrição sobre o cancro da mama em Portugal.

No capítulo 3, é feita uma apresentação da terminologia da análise de sobrevivência. Estuda-se a variável aleatória não negativa, que representa o tempo até à recidiva, ou seja, o tempo decorrido desde um instante inicial até à ocorrência de um acontecimento específico, e algumas funções vastamente utilizadas neste tipo de análise.

No capítulo 4, será feita uma análise exploratória dos dados, quer ao nível do indivíduo, quer ao nível do carcinoma.

Algumas distribuições paramétricas e não paramétricas para os tempos de vida são também apresentadas no capítulo 5, sendo dada especial atenção ao estimador não paramétrico de Kaplan-Meier [15] para estimar a curva de sobrevivência. Discute-se a comparação de distribuições de sobrevivência entre os vários grupos, fazendo-se uso do teste de log-rank.

Neste capítulo, descreve-se ainda o modelo de regressão de Cox e discutem-se alguns detalhes deste modelo de regressão semi-paramétrico, desenvolvidos especificamente para ajustamentos a dados censurados.

No capítulo 6, será feito um estudo longitudinal para dois marcadores, o marcador Antígeno Carcinogénico (CA15.3) e o antígeno carcioembriónico (CEA), em que o principal objetivo é descrever a progressão destes dois marcadores, em relação aos possíveis fatores de risco, bem como compreender como é que esses fatores influenciam a progressão do mesmo. Estes dois marcadores foram analisados separadamente, utilizando para isso, a análise prévia do modelo de análise de sobrevivência de Kaplan-Meier.

Para o estudo longitudinal, todas as covariáveis pertencentes aos tratamentos dos pacientes foram removidas. O tempo de referência utilizado foi o tempo desde o diagnóstico até a data do teste em que o evento é recorrência. Para o estudo, utilizamos o variograma empírico e teórico dos resíduos do modelo.

Por último, no capítulo 7, enumeram-se as principais conclusões do estudo realizado.

Toda a análise dos dados foi feita com auxílio do *software* R.

R é uma linguagem ambiente de desenvolvimento integrado para cálculos estatísticos e gráficos. É um software livre, que se encontra no site www.r-project.org

# Capítulo 2- Unidade de Senologia do Hospital de Braga

## 2.1. Hospital de Braga

O hospital de Braga está situado no leste da cidade de Braga, situada no norte de Portugal. A criação deste hospital permitiu alargar os cuidados médicos a utentes do distrito de Viana do Castelo e Braga.

Em 2008, foi criada a unidade de Senologia do Hospital de Braga. Este serviço opera 150 novos casos/ ano [20].

Em 2017, o hospital de Braga foi considerado o melhor do país, obtendo a classificação máxima nas cinco dimensões estudadas - excelência clínica, segurança do doente, adequação e conforto das instalações, focalização no utente e satisfação do utente.

É uma unidade hospitalar que integra o Serviço Nacional de Saúde, no âmbito de uma Parceria Público Privada celebrada através de um contrato de gestão assinado pela Administração Regional de Saúde Norte, em representação do Ministério da Saúde.

Podemos ter mais informações em https://www.hospitaldebraga.pt.

## 2.2. Cancro da mama

Segundo a Agência “International Agency for Research on Cancer”, [21] o cancro de mama é o segundo cancro mais comum a nível mundial e é o tipo de cancro mais comum em pacientes do género feminino.

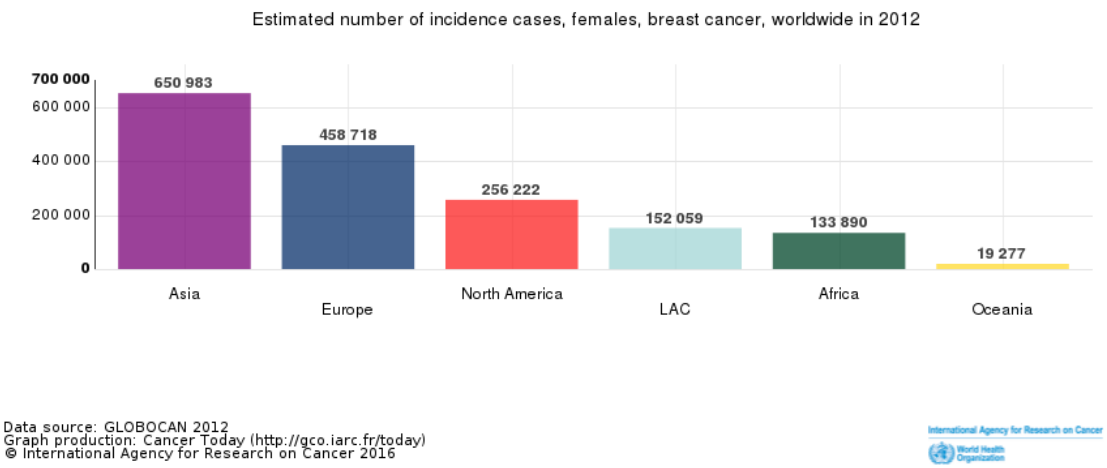



**Figura1-Estimativas do número de casos de incidência do cancro da mama em 2012 (http://globocan.iarc.fr/Default.aspx)**

Como podemos observar pela figura 1, em 2012, o cancro da mama foi mais incidente no continente asiático com 650983 casos. A Europa também tem uma taxa de incidência alta, com 458 718 caos de cancro da mama.

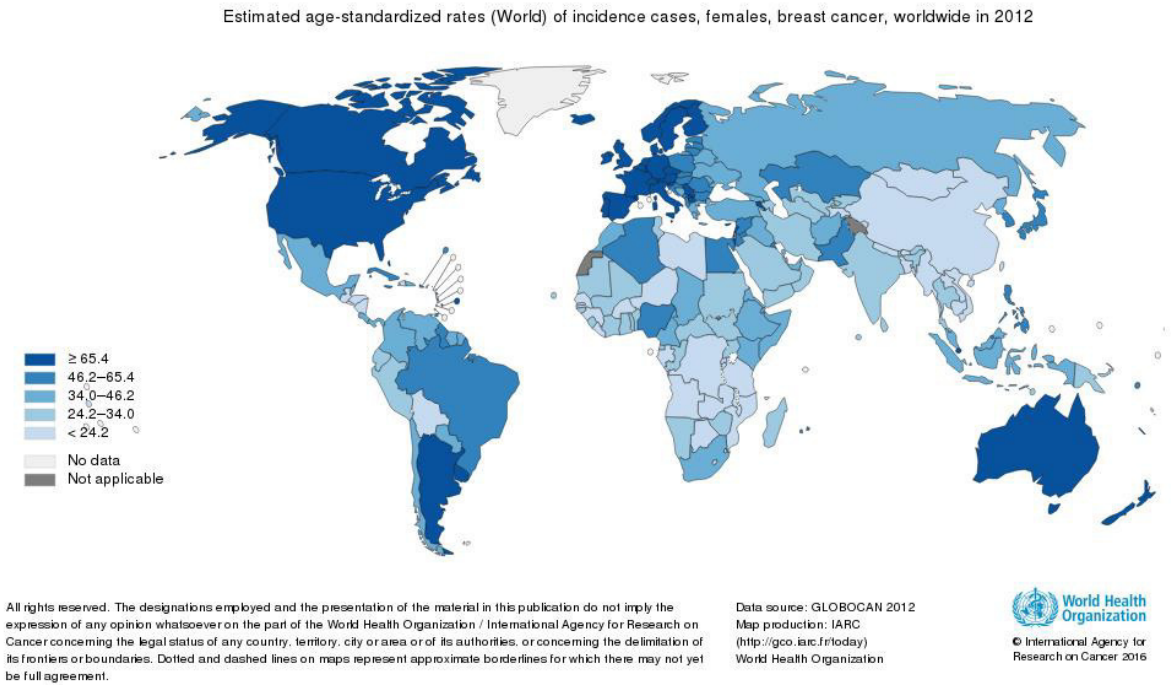



**Figura 2- Estimativas da média de idades em função da incidência do cancro da mama (http://globocan.iarc.fr/Default.aspx)**

O cancro da mama é a quinta causa mais comum de morte, com 522 000 mortes em 2012. Embora seja a causa mais comum de morte por cancro em pacientes de regiões menos desenvolvidas (com 324 000 mortes em 2012; 14,3% de todas as mortes por cancro em pacientes em regiões menos desenvolvidas), agora é a segunda causa mais comum de morte por cancro (depois do cancro do pulmão) em pacientes de regiões mais desenvolvidas (com 198 000 mortes em 2012; 15,4% de todos morte por cancro em pacientes de regiões mais desenvolvidas). A variação nas taxas de mortalidade em todas as regiões do mundo é menor do que a variação nas taxas de incidência, devido a uma melhor sobrevivência nas regiões desenvolvidas (alta incidência); em 2012, a taxa de mortalidade (por 100 000 pacientes) varia de 6 mortes na Ásia oriental, a 20 mortes na África ocidental.

Em Portugal, a taxa de incidência de cancro é alta, correspondente a 29,4%. A sua taxa de mortalidade de 16%, para o ano de 2012. [21]

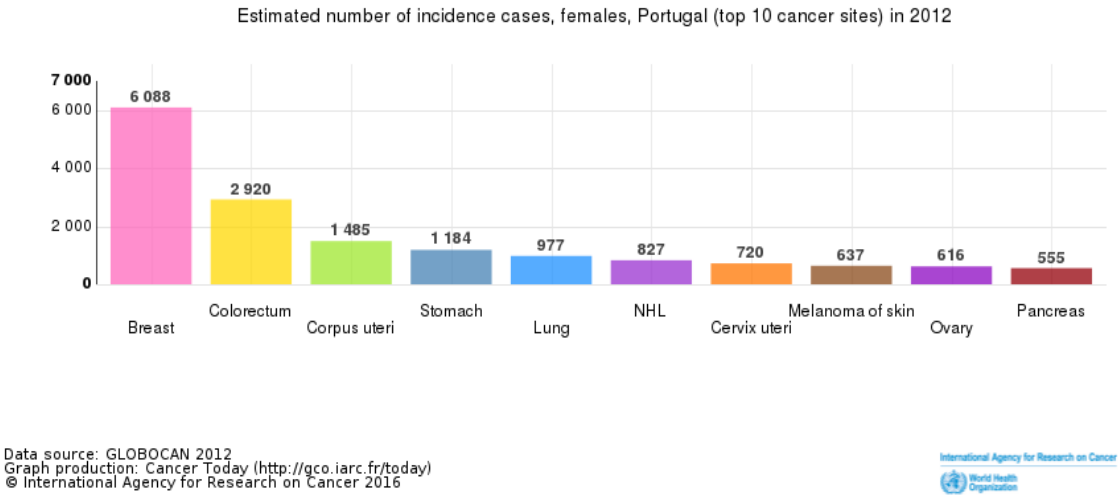



**Figura 3 - Estimativas do número de casos de incidência de cancro da mama em Portugal (http://globocan.iarc.fr/Default.aspx**

Como podemos observar pela figura 3, o cancro da mama é o tipo de cancro mais incidente no ano de 2012, e o cancro do pâncreas o menos incidente [22].

# Capítulo 3 - Metodologia

## 3.1. Análise de sobrevivência

A análise de sobrevivência é uma das áreas da estatística. Na análise de sobrevivência, a variável resposta é o tempo até à ocorrência de um evento de interesse. Este tempo é denominado tempo de falha, podendo ser por exemplo o tempo até à morte do paciente, bem como até à cura ou recidiva de uma doença.

Em estudos oncológicos, é normal o registo das datas correspondentes ao diagnóstico da doença, à remissão (após o tratamento, a paciente fica livre dos sintomas da doença), à recorrência da doença (recidiva) e à morte da paciente.

A principal característica de dados de sobrevivência é a presença de censura, que é a observação parcial da resposta. Isto refere-se a situações em que, por alguma razão, o acompanhamento do paciente foi interrompido, seja porque se perdeu o rasto do paciente, ou o paciente morreu de causas diferentes das estudadas. Isto significa que toda a informação referente à resposta se resume ao conhecimento de que o tempo de falha é superior àquele observado.

Os conjuntos de dados de sobrevivência são caracterizados pelos tempos de falha e pelos tempos de censurados. Estas duas componentes constituem a resposta. Em estudos clínicos, um conjunto de variáveis explicativas é, também, medido em cada paciente. Os seguintes elementos constituem o tempo de falha: o tempo inicial, a escala de medida e o evento de interesse (falha).

Censura

A análise de sobrevivência é a metodologia mais adequada, para analisar o tempo até um determinado acontecimento, pois permite a presença de dados censurados. Define-se então como uma observação censurada aquela que corresponde a um indivíduo cujo evento não foi possível observar, mas, em vez disso, observou-se o tempo desde a sua entrada no estudo até ao instante do abandono ou fim do estudo.

Existem vários tipos de censura, censura à esquerda, censura intervalar e censura à direita. Na censura à esquerda, o tempo de vida é inferior ao tempo durante o qual o indivíduo esteve em observação, que é o tempo de censura. A censura intervalar é quando os indivíduos são seguidos periodicamente. Censura à direita quando o tempo de vida é superior ao tempo de censura. Neste trabalho, será considerada a censura à direita.

Numa amostra de dimensão $n$, a cada indivíduo$i$ corresponde o par ( $T_i^*, \delta_i$) onde:

- $T_i^* = \mathrm{Min}(T_i, C_i)$ Designa o tempo de $follow - up$, sendo $T_i$, o tempo de vida (ou o tempo que decorreu desde a entrada em estudo até à realização do acontecimento de interesse) e $C_i$ , o tempo de censura;
- $\delta_i = I\ (\{T_i\ \leq C_i\})$ Denota-se o estado final do indivíduo, tomando o valor 1, se foi possível observar o tempo de vida $T_i$ e o valor 0, se só foi possível observar o tempo de censura $C_i$;

Truncatura

A truncatura à direita ocorre quando apenas são considerados os indivíduos para os quais se observou um determinado acontecimento de interesse durante o período em estudo.

A truncatura à esquerda é usada quando apenas são incluídos na amostra indivíduos que sobrevivem tempo suficiente para que ocorra um determinado acontecimento antes do evento de interesse.

É de realçar que o facto de existir censura não impede a existência em simultâneo de truncatura, sendo mesmo muito comum deparar com dados que são censurados à direita e truncados à esquerda.

## 3.2- Funções associadas ao tempo de vida

Seja $T$ uma variável aleatória não negativa que representa o tempo de vida de um indivíduo proveniente de uma dada população homogénea. A distribuição de $T$ pode ser univocamente especificada através de qualquer uma das seguintes

funções: função de sobrevivência $S(t)$, a função densidade de probabilidade $f(t)$, ou a função de risco$\lambda(t)$.

A função de sobrevivência e a função de risco são importantes para o estudo do tempo até à realização do acontecimento de interesse.

Sendo $F(t)$ a função de distribuição da variável aleatória $T$, esta função representa a probabilidade do tempo de vida ser inferior ou igual a um dado instante $t$:

$$F(t) = P(T \leq t), \qquad 0 \leq t \leq +\infty$$

Esta função é monótona não decrescente, contínua à direita e tal que, $F(0) = 0$ e $\lim_{t\to +\infty} F(t) = 1$ .

Quando $F(t)$ é diferenciável, define-se a função densidade de probabilidade de $T$ por:

$$F'(t) = f(t) = \lim_{t\to +\infty} \frac{P(t < T \leq t + \Delta t)}{\Delta t}$$

Dando um valor aproximado da probabilidade de ocorrência do acontecimento num intervalo] t; t +Δt ].

A função de sobrevivência $S(t)$ no instante t é dada por:

$$S(t) = P(t > t) = \int_t^{+\infty} f(u)du \ , \ 0 \leq t \leq +\infty$$

A função de sobrevivência é uma função monótona não crescente, contínua à esquerda e que satisfaz as propriedades: $S(0) = 1$ e $\lim_{t\to +\infty} S(t) = 0$ .Além disso, $f(t) = F'(t) = -S'(t)$.

A função de risco representa a taxa instantânea de ocorrência do acontecimento no instante t, e define-se por:

$$\lambda(t) = \lim_{\Delta t\to 0^+} \frac{P(t < T \leq t + \Delta t \mid T > t)}{\Delta t}$$

A função de risco satisfaz as seguintes propriedades, $\lambda(t) \geq 0$ e $\int_0^{+\infty} \lambda(\mathrm{t})dt = +\infty$, e o seu comportamento está dependente da situação que se está a analisar.

A função de risco cumulativa é definida por:

$$\Delta(t) = \int_0^t \lambda(u)du$$

As três funções anteriores relacionam-se entre si. Uma das relações mais utilizadas é:

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{d}{dt}\ln S(t)$$

dado que $S(0) = 1$, obtendo-se

$$\Delta(t) = -\ln S(t) \Leftrightarrow S(t) = \exp[-\Delta(T)]$$

## 3.3- Modelos de Sobrevivência Paramétricos

Os métodos utilizados na análise de sobrevivência podem ser: paramétrico, não paramétrico e semi-paramétrico. Os mais usuais são o semi-paramétrico e o paramétrico.

De seguida apresentam-se, algumas das distribuições mais utilizadas em análise de sobrevivência:

Seja $T$ uma variável aleatória com distribuição exponencial de parâmetro $\lambda > 0$ com função densidade de probabilidade dada por:

$f(t) = \lambda \exp(-\lambda \mathrm{t}), \quad \mathrm{t} \geq 0$ a função de sobrevivência é dada por

$S(t) = \exp(-\lambda \mathrm{t}), \quad \mathrm{t} \geq 0$ , a sua função de risco é $h(t) = \lambda, \quad \mathrm{t} \geq 0$

Como se pode observar, a função risco é constante ao longo do tempo.

Seja T uma variável aleatória com distribuição Weibull de parâmetros $\lambda$ e $\alpha$, $T \sim W(\lambda, \alpha)$. A distribuição de Weibull é definida por:

$$f(t) = \lambda\,\alpha\,t^{\alpha-1}\exp(-\lambda t^{\alpha}), \qquad t \geq 0$$

$$S(t) = \exp(-\lambda t^{\alpha}), \qquad t \geq 0$$

$$h(t) = \lambda\,\alpha t^{\alpha-1},$$

A distribuição de Weibull é uma generalização da distribuição exponencial. É mais utilizada, pois a sua função de risco tanto pode ser constante ($\alpha = 1$), monótona crescente ($\alpha > 1$) ou monótona decrescente ( $0 < \alpha < 1$).

Outra generalização da distribuição exponencial é a distribuição Gama.

Seja $T$ uma variável com distribuição Gama com parâmetros $\lambda$ e $k$. As funções densidade de probabilidade, de sobrevivência e de risco são dadas por:

$$f(t) = \frac{\lambda^{k}}{\tau\,(k)}t^{k-1}\ \exp(-\lambda t)$$

$$S(t) = 1 - I(-\lambda t, k)$$

$$h(t) = \frac{\frac{\lambda^{k}}{\tau\,(k)}t^{k-1}\ \exp(-\lambda t)}{1 - I(-\lambda t, k)}$$

onde $t \geq 0, \lambda > 0$ , $k > 0$ , $\tau\,(k)$ é a função gama, e $I(-\lambda t, k) = \frac{1}{\tau(k)}\int_0^t u^{k-1}e^{u}\,du$ é conhecida como a função gama incompleta.

A função de risco é constante quando (k = 1), monótona crescente quando (k>1) e monótona decrescente ( 0 < k < 1).

Uma variável aleatória $T$ tem distribuição log-Normal de parâmetros $\mu$ e $\sigma$ média$\mu$ e variância $\sigma^2$. A função densidade de probabilidade para $T$ é, para $t \geq 0$,

$$f(t) = \frac{1}{\sqrt{2\pi\sigma t}}\exp\left[-\frac{1}{2}\left(\frac{lnt - \mu}{\sigma}\right)^{2}\right]$$

A função de sobrevivência é dada por

$$S(t) = 1 - \emptyset\left(\frac{\ln(t) - \mu}{\sigma}\right)$$

A sua função de risco é dada por

$$h(t) = \frac{\frac{1}{\sqrt{2\pi\sigma t}} \exp\left[-\frac{1}{2}\left(\frac{\ln(t)\,\mu}{\sigma}\right)^2\right]}{1 - \emptyset\left(\frac{lnt - \mu}{\sigma}\right)}$$

onde $\mu \in R, \sigma > 0$.

## 3.4- Modelo de sobrevivência não-paramétrico

### 3.4.1- O estimador Kaplan-Meier

O estimador desenvolvido por Kaplan-Meier [15], para a função de sobrevivência também fornece informação relativa dos indivíduos em que a morte não é observada. Para se sobreviver $t$ unidade de tempo desde um instante inicial $t_0$ é preciso que se sobreviva a todos os instantes entre $t_0$ e $t_0 + t$.

A probabilidade de sobrevivência num dado instante t é dada pelo quociente entre o número de indivíduos que sobrevivem no instante t e o número de indivíduos que estão em risco imediatamente antes desse instante, admitindo que os indivíduos apresentam igual probabilidade de morte.

Considere-se uma amostra de dimensão $n$ proveniente de uma população homogénea, ou seja, em que os indivíduos não diferem quanto ao risco de morte. Nessa amostra, $m$ é o número de tempos de vida distintos tais que, $0 = t_0 < t_1 \ldots < t_m$ , com $m \leq n$. A probabilidade de sobreviver para além do instante $t_i, i = 1, \ldots, m,$ é definida como:

$$S(t_i) = P(T > t_i) = P(T > t_{i-1}) * p_i = S(t_{i-1}) * p_i = \prod_{k=1}^{i} p_k$$

Com

$$p_i = P(T > t_i | T > t_{i-1}) = \frac{P(T > t_i)}{P(T > t_{i-1})}$$

Como estamos a considerar instantes em que ocorreram mortes, a cada instante $t_i$ está associado o número de mortes ocorridas , $d_i \geq 1, i = 1, \ldots, m$ .

Para cada instante $t_i$ define-se o número$n_i$de indivíduos em risco em $t_i^-$.

O estimador de probabilidade condicional $p_i$, de sobreviver ao instante $t_i$, é dado por:

$$\hat{p}_i = \frac{n_i - d_i}{n_i} = 1 - \frac{d_i}{n_i} = 1 - \hat{q}_i \quad i = 1, \ldots, m$$

onde $\hat{q}_i$ é o estimador da probabilidade de morte em $t_i$, dada a sobrevivência até esse instante.

O estimador Kaplan-Meier para a função de sobrevivência no instante $t$ é :

$$\widehat{S_{KM}}(t) = \prod_{j:t_j \leq t} (1 - \frac{d_j}{n_j})$$

O estimador de Kaplan-Meier não permite o estudo simultâneo do efeito de diversas covariáveis no tempo de vida, nem ajustar esse efeito a eventuais combinações de outras variáveis. Nesse caso, deve usar-se um modelo de regressão, como por exemplo, o modelo de riscos proporcionais de Cox, como estudaremos a seguir.

## 3.5 -Modelo de Sobrevivência Cox

Na análise de sobrevivência, muitas vezes existem covariáveis que podem estar relacionadas com o tempo de sobrevivência.

Essas covariáveis devem ser incluídas na explicação do possível efeito no tempo de sobrevivência. Uma das alternativas é a introdução de covariáveis no modelo de riscos proporcionais. Este problema pode ser abordado através de um estudo de um modelo de regressão que tenha em consideração a ocorrência de observações censuradas, sendo uma possível escolha, o modelo de riscos proporcionais, introduzido por Cox [6]. A importância deste modelo deve-se ao facto de ser um modelo semi-paramétrico.

No modelo de Cox a função de risco é dada por:

$$h(t; z) = h_0(t)e^{\beta^T Z}$$

onde $h_0(t)$ é uma função de risco não negativa e$\beta$ um vetor de coeficientes de regressão, e $Z$ um vetor de covariáveis.

No modelo existe uma componente multiplicativa não paramétrica ($h_0(t)$) e uma outra paramétrica $e^{\beta^T Z}$, sendo esta a razão do modelo ser designado por semi-paramétrico.

Assim, duas características deste modelo são:

- As funções de risco correspondentes a dois indivíduos diferentes com vetores de covariáveis $Z_1$ e $Z_2$ são proporcionais logo a razão entre estas duas funções em qualquer instante $t$:

$$\frac{h(t,Z_1)}{h(t,Z_2)} = \frac{h_0(t)e^{\beta^T Z_1}}{h_0(t)e^{\beta^T Z_2}} = e^{[\beta^T(Z_1 - Z_2)]}$$

- Não depende de $t$;

- As covariáveis afetam a função de risco proporcionais, de modo multiplicativo de acordo com o fator $e^{\beta^T Z}$, que é designado por risco relativo.

O modelo de regressão de Cox é caracterizado pelos coeficientes $\beta$, que medem o efeito das covariáveis sobre a função de risco. Cox [6], construiu uma função de verosimilhança que não depende de $h_0(t)$, permitindo, a realização de inferência sobre $\beta$ sem que seja necessário especificar $h_0(t)$.

Considerando $n$, indivíduos em estudo, e $k$, instantes distintos de observação de eventos, tal que $t_1 < t_2 < \cdots < t_k$, em que, $k \leq n$, o risco no instante $t_i$, designado por $R_i$ é definido por:

$$R_i = R(t_i) = \{j : t_j \geq t_i\}.$$

Os índices associados aos indivíduos surgem imediatamente antes do instante $t_i$.

A função de verosimilhança, proposta por Cox [6] para a realização de inferência sobre $\beta$ é dada por:

$$L(\beta) = \prod_{i=1}^{k} \frac{e^{(\beta^T z_i)}}{\sum_{l \in R_i} e^{(\beta^T z_l)}}$$

Em que $z_i$ é o vetor de variáveis explicativas associadas ao indivíduo com evento no instante $t_i$. Cox [8] designou a função $L(\beta)$ por função de verosimilhança parcial.

Embora não se trate de uma função, uma vez que não permite a obtenção de um estimador do vetor de parâmetros $\beta$, no entanto, verifica as propriedades dos estimadores de máxima verosimilhança.

A função de verosimilhança proposta por Cox [7] não depende de $h_0(t)$, o que permite a realização de inferência sobre o vetor de parâmetros $\beta$, sem que seja necessário fazer restrições à forma de $h_0(t)$.

Este método tem a presença de uma componente não-paramétrica $h_0(t)$, na função de verosimilhança, o que o torna inapropriado. Assim, Cox [8], propôs o método de máxima verosimilhança parcial que condiciona a verosimilhança e, desta forma, elimina a função de base desconhecida. Tal como Cox [8] mostrou, o método de construção de verosimilhança parcial, leva às propriedades assintóticas de inferência baseada na verosimilhança.

Nas situações em que simultaneamente tenham ocorrido falhas em mais do que um individuo, dando origem a valores iguais, a função de verosimilhança parcial não poderá ser aplicada.

Suponhamos que $n$ indivíduos foram observados nos instantes $t_1 < t_2 < \cdots < t_k$. Seja $d_i$o número de eventos ocorridos no instante $t_i$ e $Z_{ij}$ o vetor de covariáveis explicativas associadas ao indivíduo $j$, cujo evento ocorre em $t_i$, $j = 1, \dots, d_i$ , $i = 1, \dots, k$. Se, o número de $d_i$ com evento em $t_i$ for pequeno, então a função de verosimilhança parcial pode ser aproximada pela função, proposta por [2],

$$L(\beta) = \prod_{i=1}^{k} \frac{e^{(\beta^T s_i)}}{\left[\sum_{l \in R_i} e^{(\beta^T z_l)}\right]^{d_i}}$$

onde $s_i = \sum_{j=1}^{d_i} Z_{ij}$, para $i = 1, \dots, k$.

### 3.5.1. Seleção de modelos de Cox e verificação de pressupostos

A escolha do modelo de regressão de Coxé difícil pois raramente existe um único melhor modelo. Por conseguinte, é necessário um modelo que proporcione um

bom ajuste aos dados. Normalmente recorre-se a procedimentos de busca sequencial.

Os métodos de busca sequencial têm em comum a abordagem geral de estimar a equação de regressão com um conjunto de variáveis e então acrescentar seletivamente, ou eliminar variáveis, maximizando a previsão com o menor número de variáveis empregues [14]. A abordagem sequencial mais comum para a seleção de variáveis é a estimação *stepwise*. Os procedimentos *stepwise* dividem-se em *forward selection*, método composto e *backward elimination* [17].

O critério de informação de Akaike [1], é outro possível critério de seleção de um modelo, que se baseia na função log-verosimilhança com a introdução de um fator de correção como modelo de penalização da complexidade do modelo, examina a seguinte estatística:

$$AIC = -2loglikelihood + 2 * r$$

Onde $r$ é o número de parâmetros do modelo. Um valor baixo para AIC é considerado como representativo de um melhor ajustamento e, assim, na seleção de modelos deve-se ter como objetivo a Minimização de AIC.

Existem vários métodos para verificar a validade de riscos proporcionais.

Esse pressuposto pode ser validado através de um gráfico $\log - log\ S(t)$ versus $t$. As curvas obtidas devem ser paralelas. Assim, o modelo de riscos proporcionais é inadequado quando as curvas se intersetam.

Outra aproximação que pode ser utilizada é o gráfico dos resíduos de Schoenfiels, propostos por Schoenfiels [19], que são muito úteis na avaliação da hipótese de riscos proporcionais. A cada indivíduo não corresponde apenas um resíduo mas um conjunto de valores, onde cada valor é referente a cada uma das variáveis explicativas incluídas no modelo de regressão de Cox.

## 3.6- Modelos de dados longitudinais

Dados longitudinais, são observações em que os indivíduos são medidos repetidas vezes ao longo do tempo em relação a uma mesma característica, num ou mais grupos de tratamento, sendo o próprio tempo um factor de interesse [12].

Os dados longitudinais podem ser obtidos de uma forma prospectiva ou retrospetiva; na primeira situação os indivíduos são seguidos ao longo do tempo, na segunda situação múltiplas medições em cada indivíduo são extraídas do seu historial.

O principal objetivo de um estudo longitudinal é caracterizar as alterações da variável resposta com o tempo assim como, determinar se essas alterações se relacionam com um conjunto de covariáveis, isto é, com o conjunto de fatores previamente escolhidos, que não o tempo.

Vão-se seguidamente indicar algumas das características dos estudos longitudinais:

- Cada indivíduo, tem um vetor resposta constituído pelas medições repetidas que estão provavelmente correlacionadas e, como tal, a estrutura de correlação desempenha um papel importante na estimação dos parâmetros do modelo a ajustar aos dados;
- Existir uma elevada variabilidade entre os diferentes indivíduos;
- Os indivíduos terem diferente número de observações e de estas terem sido feitas em ocasiões distintas;
- Em alguns estudos longitudinais poder acontecer que não só a variável resposta se altera ao longo do tempo mas também o valor das covariáveis.

### 3.6.1 Regressão linear

A análise de regressão tem como objetivo verificar a existência de uma relação linear entre uma variável dependente com uma ou mais variáveis independentes.

O comportamento da variável dependente em relação à variável independente pode apresentar-se de várias maneiras: linear, quadrática, cúbica, exponencial, entre outras. O diagrama de dispersão é muitas vezes utilizado, contudo os pontos podem não se ajustar na perfeição à curva do modelo matemático proposto.

Um dos métodos que se pode utilizar é o método dos mínimos quadrados (OLS), que é a soma dos quadrados das distâncias entre os pontos do diagrama e os respetivos pontos na curva da equação, obtendo-se, desta forma uma relação entre $X$ e $Y$ com o menor erro possível.

Modelo de regressão linear simples pode traduzir como

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

em que

$Y_i$ é o valor observado para a variável dependente

$\beta_0$ é a constante de regressão, que representa o intercepto da reta com o eixo dos Y

$\beta_1$ é o coeficiente de regressão, que é a variação de Y em função da variação de X.

$X_i$ i-ésimo nível da variável independente

$e_i$ é o erro que esta associado à distância entre o valor observado e o correspondente ponto na curva.

Para se obter a equação estimada, vamos utilizar o OLS, visando a Minimização dos erros. Assim temos:

$e_i = Y_i - \beta_0 - \beta_1 X_i$ elevando ao quadrado

$e_i{}^2 = [Y_i - \beta_0 - \beta_1 X_i]^2$ aplicando o somatório

$$\sum_{i=1}^{n} e_i{}^2 = \sum_{i=1}^{n} [Y_i - \beta_0 - \beta_1 X_i]^2$$

Por meio da obtenção de estimadores $\beta_0$ e $\beta_1$, que minimizam o valor obtido na expressão anterior é possível alcançar o mínimo da soma dos quadrados dos erros.

Para encontrar o mínimo derivamos a função em relação à variável de interesse e iguala-la a zero, o que obtemos:

$$\widehat{\beta_1} = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum {x_i}^2 - \frac{(\sum x_i)^2}{n}}$$

e

$$\widehat{\beta_0} = \bar{Y} - \widehat{\beta_1}\bar{X}$$

Uma vez obtidas estas estimativas, podemos escrever a equação estimada:

$$\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} X$$

### 3.6.2. Terminologia e notação

Designa-se por $Y_{ij}$ $(j = 1, \dots, m_i)$ o valor da resposta na ocasião j para o indivíduo $i$ , $(i = 1, \dots, n)$, e pode ser descrita como:

$$Y_{ij} = \mu_{ij} + \varepsilon_{ij}$$

Para cada componente $Y_{ij}$ tem-se $E(Y_{ij}) = \mu_{ij}$

O conjunto de todas as variáveis resposta é dado por $N = \sum_{i=1}^{n} m_i$ onde n é o número total de medições do conjunto de dados. Se $n = m_i$ significa que estamos perante um estudo balanceado, isto é, todos os pacientes foram medidos nos mesmos tempos. Caso $m_i \neq n$ então o estudo era não balanceado.

Diggle [10] considera três tipos de modelos para análise de dados longitudinais: o modelo marginal, o modelo de efeitos aleatórios e o modelo de transição.

O modelo marginal tem como objetivo fazer inferência sobre o valor médio populacional e na sua dependência com as covariáveis, não necessitando de

assumir pressupostos para a distribuição da variável resposta, apenas de um modelo de regressão.

Neste modelo o valor esperado marginal $E(Y_{ij})$, é modelado como função das covariáveis, não sendo condicionado a outras variáveis resposta ou a efeitos aleatórios não observáveis [11]. Como as medições repetidas, em cada indivíduo, não têm tendência a ser independentes, a análise marginal tem que incluir pressupostos em relação à correlação.

O modelo marginal tem a vantagem do valor esperado da variável resposta e a covariância serem modelados separadamente [10].

Os modelos com efeitos aleatórios são utilizados para descrever as alterações da resposta média de cada indivíduo e a relação destas alterações com as covariáveis.

Estes modelos permitem a acomodação de respostas quer sigam a distribuição Gaussiana ou não. Para além, disso podem modelar a sobredispersão e a correlação através da incorporação de efeitos aleatórios.

Nos modelos de transição, o valor esperado e a dependência temporal são simultaneamente modelados, condicionando uma resposta a outras respostas ou a um subconjunto particular de outras respostas.

Apresentam-se em seguida, modelos para a análise de dados longitudinais dependendo da natureza da variável resposta [13]:

- Modelo misto
- Modelo de equação de estimação generalizada (GEE)
- Modelos bayesianos
- Modelos não paramétricos/semiparamétricos.

O modelo mais utilizado é o modelo de equações de estimação generalizada (GEE) [16], pois permite a modelação do valor médio da variável resposta em função das covariáveis com a acomodação da dependência entre as observações do mesmo indivíduo sem necessitar de especificar a distribuição conjunta do vetor associado e esse indivíduo.

O modelo de transição é um caso particular dos modelos onde o valor esperado e a dependência temporal são simultaneamente modelados, condicionando uma resposta a outras respostas ou a um subconjunto particular de outras respostas.

Neste modelo a correlação entre $Y_{i1}, \dots, Y_{im_i}$, existe porque os valores passados de e $Y_{i1}, \dots, Y_{im_i-1}$, influenciam o valor observado de$Y_{im_i}$. As variáveis anteriores a $Y_{iTi}$ são tratadas como covariáveis adicionais [13]. Estes modelos são úteis para modelar o valor esperado da resposta condicionada quer às covariáveis, quer às observações passadas.

O modelo de regressão de mínimos quadrados (OLS) assume independência entre quaisquer duas medições do mesmo sujeito ou sujeitos diferentes. No entanto, em determinados estudos torna-se necessário considerar outros modelos que estudem a variabilidade entre indivíduos e dentro do próprio indivíduo, tais como os modelos longitudinais.

Os modelos longitudinais propõem a decomposição da variabilidade $\varepsilon_{it}$ na variabilidade entre indivíduos e dentro do próprio indivíduo, ao longo do tempo.

O modelo longitudinal é descrito [13]:

$$Y_{ij} = \mu_{ij} + d'_{ij}U_i + w_i(t_{ij}) + Z_{ij}$$

Onde $U_i$ são $n$ realizações i.i.d de $MVN(0, \vartheta^2)$ representando os efeitos aleatórios ao nível do indivíduo, e $d'_{ij}$ é o vetor de covariáveis para o efeito aleatório;

$w_i(t_{ij})$ representa um processo de tempo contínuo gaussiano em que $E\left(w_i(t_{ij})\right) = 0$ *e* $Var\left(w_i(t_{ij})\right) = \vartheta^2$, representa a variabilidade entre os sujeitos. A correlação entre duas medições de um indivíduo é descrita como $Corr\,((w_i(t_{ij}), w_i(t_{ik})) = \rho\,(t_{ij}, t_{ik})$;

Finalmente, $Z_{ij}$ são as $n$ realizações i.i.d, $N(0, \tilde{\imath}^2)$ e representam a variabilidade não explicada ou erro de medição.

Uma vez que, $w_i(t_{ij})$ é considerado um processo estacionário temos $\rho\left(t_{ij}, t_{ik}\right) = \rho\,(|t_{ij}, t_{ik}|)$ .

Podemos ter diferentes definições da função $\rho\,(|t_{ij}, t_{ik}|)$.

Se considerarmos a correlação entre $w_i(t)$ e $w_i(t-u)$, determinada pela função de autocorrelação $\rho(u) = \exp(-\frac{1}{\emptyset}|u|)$, teremos um modelo longitudinal que explica uma estrutura de correlação exponencial entre indivíduos. Um modelo longitudinal com $\rho(u) = \exp(-\frac{1}{\emptyset}u^2)$, explica uma estrutura de correlação gaussiana entre indivíduos.

Onde $\emptyset$ é o parâmetro de referência que determina o grau onde a correlação estabiliza.

Para modelar a parte fixa do modelo longitudinal, $\mu_{ij}$ podemos considerar o modelo com um ponto de mudança δ no efeito do tempo na progressão média da variável resposta.

O ponto de mudança é o momento em que há uma alteração no declive da progressão média da variável resposta. Considerando δ o ponto de mudança, temos $E(Y_{ij}) = \mu_{ij}$ com

$$\mu_{ij} = \begin{cases} X_{ij}\,\beta + \propto_1 t_{ij}\; se\; t_{ij} < \delta \\ X_{ij}\,\beta + \propto_2 (t_{ij} - \delta)\; se\; t_{ij} \geq \delta \end{cases}$$

Onde $X_{ij}$ representa o vetor de covariáveis, $\beta$ o vetor de coeficientes de regressão desconhecido, $\propto_1$e $\propto_2$ os coeficientes que representam o declive de uma curva antes e após o ponto de mudança, respetivamente.

Consideremos a série completa de $N$ medições, Y como realização de um vetor gaussiano aleatório multivariado com $Y \sim MVN(X\beta, \eta^2 V)$ onde $X$ é uma matriz $N * p$ de todos os valores das variáveis explicativas ρ e $\eta^2 V$, com uma matriz bloco-diagonal com entradas não-nulas. Cada um representando a matriz de variação para o vetor de medidas resultantes de um único indivíduo.

Para a estimativa do parâmetro adotamos o método da máxima verosimilhança cuja função de verosimilhança do logaritmo associada, para os dados y observados, sob a distribuição Gaussiana, é dada por:

$$L(\beta, \eta^2, V_0) = -\frac{1}{2}\{nmlog(\eta^2) + mlog(|V_0|) + \frac{1}{\eta^2}(y - X\beta)' V^{-1}(y - X\beta)\}$$

A estimativa de máxima verosimilhança para $\beta$, em que $V_0$ é o estimador obtido pelo método dos mínimos quadrados [10] é dado por:

$$\hat{\beta}(V_0) = (X'V^{-1}X)^{-1}X'V^{-1}y$$

Substituindo na equação anterior obtemos

$$L(\hat{\beta}(V_0),\, \eta^2\,, V_0) = -\frac{1}{2}\{nmlog(\,\eta^2) + mlog\,(|\,V_0|) \,+\, \frac{1}{\eta^2}\, RSS(V_0)\}$$

Onde $RSS(V_0) = \,(\,y - X\,\hat{\beta}(V_0\,))'V^{-1}\,(y - X\hat{\beta}(V_0))$.

O estimador de máxima verosimilhança para $\eta^2$, com $V_0$fixo, é obtido pela diferenciação de $L(\hat{\beta}(V_0),\, \eta^2\,, V_0) = -\frac{1}{2}\{nmlog(\,\eta^2) + mlog\,(|\,V_0|) \,+\, \frac{1}{\eta^2}\, RSS(V_0)\}$

em ordem a $\eta^2$, obtendo-se:

$$\widehat{\eta^2}(V_0) = \frac{RSS(V_0)}{nm}$$

A função de log-verosimilhança para $V_0$ a parte do termo constante, é obtido substituindo $\hat{\beta}(V_0) = (X'V^{-1}X)^{-1}X'V^{-1}y$ em $\eta^2(V_0) = \frac{RSS(V_0)}{nm}$ obtemos:

$$L_r(V_0) = -\frac{1}{2}m\{\,nlogRSS(V_0) + \,log(V_0)\}$$

Maximizando a função anterior obtemos o estimador de máxima verosimilhança para $V_0$. Substituindo $\hat{V}_0$ em:

$$\hat{\beta}(V_0) = (X'V^{-1}X)^{-1}X'V^{-1}y$$

e

$$\widehat{\eta^2}(V_0) = \frac{RSS(V_0)}{nm}$$

Obtemos os estimadores de máxima verosimilhança para $\hat{\beta}$ e $\widehat{\eta^2}$, respetivamente.

### 3.6.3. Dados omissos

Uma das características dos dados longitudinais é a diferença do número de observações entre os indivíduos e de estas terem sido feitas em ocasiões distintas.

Esta característica leva à existência de valores omissos, em que, há perda de informação e redução na precisão com a qual as alterações na resposta média ao longo do tempo podem ser estimadas.

### 3.6.4. Variograma

As estruturas de correlação espacial são geralmente representadas através do seu variograma em vez da estrutura de correlação [9].

Para modelar a estrutura de correlação para cada modelo analisamos o variograma empírico de resíduos do modelo OLS saturado para a média da variável resposta.

O variograma de um processo $Y(t)$ estocástico é dado por:

$$V(u) = \frac{1}{2} var\{Y(t) - Y(t-u)\}, u \geq 0$$

Para um processo estacionário, a função de autocorrelação$\rho(u)$ e a variação de $Y(t)$, $\sigma^2$, estão relacionados por:

$$\rho(u) = \sigma^2\{1 - \rho(u)\}$$

A estimativa do variograma empírico está baseada no cálculo do quadrado das diferenças entre pares de resíduos $v_{ij} = \frac{1}{2}(r_{ij} - r_{ik})^2$, e as correspondentes diferenças temporais $u_{ijk} = t_{ij} - t_{ik}$ onde $r_{ij} = Y_{ij} - u_{ij}$, e $j < k = 1, \dots, m_i$

A função de autocorrelação em qualquer intervalo de tempo $u$ é estimada pelo variograma dado por:

$$\hat{\rho}(u) = 1 - \frac{\hat{\gamma}(u)}{\hat{\sigma}^2}$$

Onde $\hat{\gamma}(u)$ é a média de todos os $v_{ij}$ , correspondente aquele valor especifico de $u$ e $\hat{\sigma}^2$ é a variância estimada do processo.

# Capítulo 4- Análise exploratória

Os dados analisados são relativos a pacientes do Hospital de Braga a quem foi diagnosticado cancro da mama, entre os anos 2008 e 2013. Neste estudo, também se encontram dados referentes a pacientes que já tinham sido diagnosticados com cancro da mama anteriormente a esta data e continuavam em consultas de seguimento.

Recolheram-se informações de 596 pacientes, dos quais 56 pacientes foram excluídos por obedecerem, pelo menos, a um dos seguintes critérios:

- Ser do género masculino;
- Não haver informação acerca do diagnóstico ou tratamento da mesma;
- Neoplasia benigna da mama.

Das 540 pacientes, 19 apresentaram cancro bilateral que foram tratados como casos independentes por aconselhamento médico, o que perfaz um total de 559 casos em estudo.

A Tabela 1 tem a listagem de todas as variáveis retiradas dos registos médicos dos doentes ao nível do indivíduo, enquanto a Tabela 2 regista as variáveis ao nível do carcinoma. É dado um pequeno resumo de cada uma das variáveis.

## 4.1- Variáveis explicativas ao nível do indivíduo

**Tabela 1-Variáveis explicativas ao nível do indivíduo**

| Variável | Descrição | Categoria |
|---|---|---|
| Data de nascimento | Data de nascimento | Data |
| Distrito | Distrito a que pertence a paciente | Categórica |
| Concelho | Concelho a que pertence a paciente | Categórica |
| Freguesia | Freguesia a que pertence a paciente | Categórica |
| Estado Civil | Estado civil da paciente | Categórica |
| Profissão | Profissão da paciente | Categórica |
| Habilitações literárias | Escolaridade da paciente | Categórica |
| Menarca | Idade da menarca | Numérica |
| Paridade | A paciente teve ou não filhos | Dicotómica |
| Nº de partos | Número de filhos que a paciente teve | Numérica |
| Amamentação | A paciente amamentou ou não | Categórica |
| Duração da amamentação | Duração da amamentação | Numérica |
| Menopausa | Encontra-se em menopausa ou não | Categórica |
| Menopausa cirúrgica | A menopausa foi cirúrgica ou não | Categórica |
| Idade à menopausa | Idade da paciente no início da menopausa | Numérica |
| THS | Tratamento Hormonal de substituição | Categórica |
| ACO | Toma do anticoncecional oral | Categórica |
| Historial familiar | Historial familiar com cancro | Categórica |
| Grau de parentesco | Qual o parentesco do familiar | Categórica |
| Mamografia anterior à consulta | Mamografia anterior à consulta de diagnóstico | Categórica |
| Data da mamografia | Data da mamografia feita | Data |
| Ecografia anterior à consulta | Ecografia anterior à consulta de diagnóstico | Categórica |
| Data da ecografia | Data da ecografia feita | Data |
| Data da 1ª consulta | Data da 1ª consulta no Hospital de Braga | Data |
| Idade ao diagnóstico | Idade ao diagnóstico | Numérica |
| Idade ao 1º parto | Idade ao primeiro parto | Numérica |
| Estado. Vital | Estado Vital | Categórica |

## 4.2- Variáveis explicativas ao nível do carcinoma

**Tabela 2 - Variáveis explicativas ao nível do carcinoma**

| Variável | Descrição | Categoria |
|---|---|---|
| Tipo de carcinoma | Tipo de carcinoma que a paciente teve | Categórica |
| Grau de diferenciação | Avaliação do grau do carcinoma | Categórica |
| Presença de Carcinoma associado | Presença de Carcinoma associado | Categórica |
| Tipo de carcinoma associado | Tipo de carcinoma associado | Categórica |
| Imagens de Invasão vascular venosa | Imagens de Invasão linfática venosa | Categórica |
| Imagens de invasão vascular linfática | Imagens de invasão vascular linfática | Categórica |
| RE | Recetores de estrogénio | Categórica |
| RP | Recetores de progesterona | Categórica |
| HER-2/neu | Receptor transmembranar | Categórica |
| Índice proliferativo ki-67 | Anticorpo monoclonal | Categórica |
| Tratamento Neoadjuvante | Paciente fez tratamento Neoadjuvante | Categórica |
| Cirurgia | Saber se a paciente fez cirurgia | Categórica |
| Tipo de cirurgia | Tipo de cirurgia feita pela paciente | Categórica |
| Esvaziamento axilar | Saber se foi feito esvaziamento axilar | Categórica |
| Tipo de recidiva | Tipo de recidiva que a paciente teve | Categórica |
| Pesquisa de gânglios sentinela | Se foi feita uma pesquisa de gânglios | Categórica |
| Tratamento Primário | Tipo de tratamento que a paciente teve | Categórica |
| Tumor primário | Tipo de tumor primário que a paciente teve | Categórica |
| Classificação patológica | Classificação patológica do tumor | Categórica |
| Metastases à distância | Classificação da metástase | Categórica |
| Padrão arquitetural do carcinoma | Padrão arquitetural do carcinoma | Categórica |
| Tamanho máximo | Tamanho máximo do carcinoma | Categórica |
| Grau nuclear do carcinoma associado | Grau do carcinoma associado | Categórica |
| Resultado gânglio sentinela | Resultado dos gânglios de sentinela | Categórica |
| TP | Tamanho do tumor ao diagnóstico | Categórica |
| Grau | Grau de diferenciação do carcinoma | Categórica |
| Gânglios regionais | Gânglios regionais | Categórica |
| Estadiamento | Estadiamento ganglionar | Categórica |
| Índice de Nottingham | Sistema de classificação de Nottingham | Categórica |
| QT | Quimioterapia | Categórica |
| RT | Radioterapia | Categórica |
| HT | Hormonoterapia | Categórica |

Esta base de dados inclui todas as pacientes que estavam em tratamento à data da criação da unidade de senologia, 1 de janeiro de 2008 (isto é, diagnóstico é anterior a esta data), e ainda todas as pacientes diagnosticadas, com primeira consulta no Hospital de Braga, entre os anos 2008 e 2012. A tabela 3 indica a evolução do número de casos ao longo dos 5 anos de estudo.

**Tabela 3 - Evolução do número de casos de cancro da mama no hospital de Braga**

| Total = 559 | Ano de diagnóstico | Número | % |
|---|---|---|---|
| | Anteriormente | 175 | 31 |
| | 2008 | 96 | 17 |
| | 2009 | 138 | 24,68 |
| | 2010 | 110 | 19,68 |
| | 2011 | 35 | 6,26 |
| | 2012 | 4 | 0,7 |
| | 2013 | 1 | 0,17 |

A tabela 4, indica as estimativas relativas ao tempo de tratamento das pacientes com cancro da mama.

**Tabela 4 - Estimativas das características clínicas relativas ao tratamento das pacientes com cancro da mama**

| | Antes de 2008 | | Depois de 2008 | |
|---|---|---|---|---|
| Variáveis | Número de casos | % | Número de casos | % |
| Intervalo entre o diagnóstico E o tratamento | | | | |
| Menos de 1 mês | 37 | 19,8 | 49 | 13,2 |
| Até 1 mês | *79* | 42,2 | 174 | 46,8 |
| Mais de 1 mês e menos de 3 | *45* | 24,1 | 84 | 22,6 |
| Mais de 3 meses e menos de 6 | *12* | 6,4 | 43 | 11,6 |
| Mais de 6 e menos de 9 meses | 3 | 1,6 | 9 | 2,4 |
| Mais de 9 meses e menos de 12 | 11 | 5,9 | 13 | 3,5 |
| Total | 187 | 100% | 372 | 100% |

A média das idades é de 59 anos e o desvio padrão é de 14,24 anos. A idade varia entre um mínimo de 20 e o máximo de 92 anos.

**Tabela 5- Medidas de localização e dispersão de variáveis quantitativas**

| Variável | Média | Min | Máx | Med | Desvio padrão | Quartis | | | NA'S |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | 25% | 50% | 75% | |
| Idade da menarca | 13,51 | 9 | 19 | 13 | 1,87 | 12 | 13 | 15 | 238 |
| Idade da menopausa | 48,63 | 33 | 58 | 50 | 5,05 | 45 | 50 | 52 | 371 |
| Idade ao primeiro parto | 25,3 | 15 | 40 | 25 | 4,65 | 22 | 25 | 28 | 335 |
| Duração da amamentação | 9,54 | 1 | 72 | 6 | 10,09 | 2 | 6 | 12 | 365 |
| Idade ao diagnóstico | 58,69 | 20 | 92 | 58 | 14,24 | 47 | 58 | 71 | - |

Das pacientes que foram observadas, 209 amamentaram, 14 não amamentaram, e não se obteve a informação de 336 pacientes. A média do tempo de amamentação foi de 9,54 meses, tal como podemos verificar na Tabela 5.

Na tabela 6, podemos verificar a distribuição das idades das pacientes ao diagnóstico de acordo com a faixa etária.

**Tabela 6- Distribuição de localização das idades ao diagnóstico de acordo com a faixa etária**

| Intervalo | Número de pacientes | Percentagem |
|---|---|---|
| 15-44 | 95 | 17% |
| 44-54 | 146 | 26% |
| 55-64 | 125 | 22,4% |
| 65-74 | 96 | 17,2% |
| >74 | 97 | 17,4% |
| Total | 559 | 100% |

O grupo que teve um maior número de pacientes diagnosticadas com cancro da mama está compreendido entre as idades 44 e 63 anos, correspondendo a uma percentagem de 26%, como verificado na Tabela 6.

É também importante verificar a distribuição do número de partos, como podemos verificar pela tabela 7.

**Tabela 7- Distribuição do número de partos por paciente**

| Número de partos | Quantidade | Proporção |
|---|---|---|
| 0 | 41 | 11,6% |
| 1 | 65 | 18,4% |
| 2 | 132 | 37,3% |
| 3 | 52 | 14,7% |
| 4 | 22 | 6,2% |
| >4 | 42 | 11,9% |
| NA'S | 205 | 36,7% |
| Total | 100 | |

Como podemos verificar pela Tabela 7, as pacientes tiveram, na sua maioria, dois partos, correspondendo a uma percentagem de 37,3%.

A tabela 8, mostra-nos o número de pacientes por distrito.

**Tabela 8- Número de pacientes por distrito**

| Distrito | Número de pessoas | Proporção |
|---|---|---|
| Braga | 539 | 96% |
| Porto | 4 | 0,7% |
| Viana do Castelo | 16 | 2,9% |

Podemos verificar que a maioria das pacientes é do distrito de Braga.

A média da idade ao diagnóstico não difere muito relativamente ao distrito das pacientes, como podemos observar pela tabela 9:

**Tabela 9 - Idade ao diagnóstico por distrito**

| Distritos<br>Idade ao diagnóstico | Braga | Porto | Viana do Castelo |
|---|---|---|---|
| Média | 59 | 44 | 55 |
| Desvio-padrão | 14,23 | 8,9 | 14,09 |
| Mínimo | 20 | 31 | 42 |
| Máximo | 92 | 51 | 81 |

Quanto à escolaridade, observou-se que o nível mais frequente foi do 4º ano, conforme se pode ver na figura 4. Não responderam 129 pessoas.

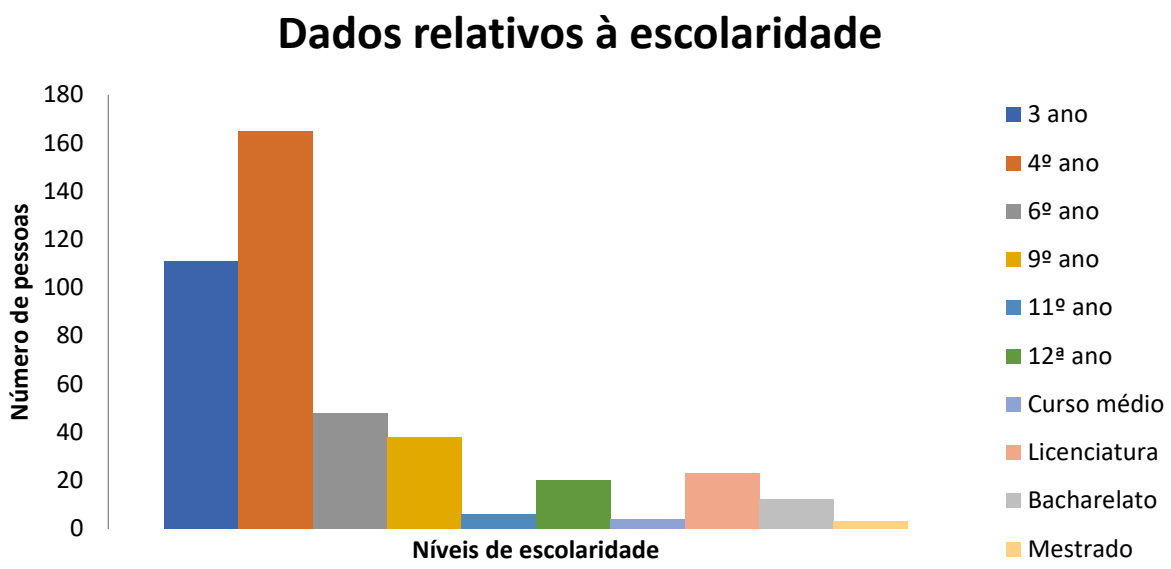



**Figura 4- Dados relativos à escolaridade**

Relativamente ao estado civil, o mais frequente é casada com 73%, sendo o menos frequente o estado "Divorciada" com 0,39%.

O número de pacientes que tiveram menopausa foi de 343, o que corresponde a uma percentagem de 63,8%, enquanto 195 não tinham menopausa, o que corresponde a uma percentagem de 36,2%, não se obteve a informação de 21pacientes correspondendo a (3,8%). Das pacientes que tiveram menopausa, 27 foi uma menopausa cirúrgica (9,2%), enquanto 265 (90,8%) foi uma menopausa não cirúrgica; 267 não responderam (47,8%).

A média das idades das pacientes que tiveram menopausa cirúrgica foi de 44,7 anos.

**Tabela 10- Dados descritivos do tipo de menopausa**

| | Tipo de menopausa | | NA'S (%) | Total |
|---|---|---|---|---|
| | Cirúrgica (%) | Não cirúrgica (%) | | |
| Quantidade | 27 (9,2%) | 265 (90,8) | 267(47,8) | 559 |
| Média das idades | 44,7 | 49,3 | | |
| Desvio-padrão | 6,4 | 4,5 | | |
| Mínimo | 33 | 38 | | |
| Máximo | 57 | 58 | | |

Como podemos observar pela tabela 11, também não há diferenças acentuadas entre os dados descritivos para a idade na menopausa, comparativamente a idade ao diagnóstico.

| Idade ao diagnóstico / Idade à menopausa | 15-44 | 44-54 | 55-64 | 65-74 | >74 |
|---|---|---|---|---|---|
| Média | 41 | 46 | 50 | 48 | 49 |
| Desvio-padrão | NA | 4,18 | 5,17 | 5,03 | 4,73 |
| Mínimo | 41 | 34 | 33 | 38 | 38 |
| Máximo | 41 | 51 | 58 | 57 | 57 |

**Tabela 11- Dados descritivos da idade ao diagnóstico e da idade à menopausa**

Como poderemos verificar pela figura 5, a idade ao diagnóstico não está correlacionada com a idade da menopausa, em que o valor da correlação= 0,060 e o valor da prova 0,4167.

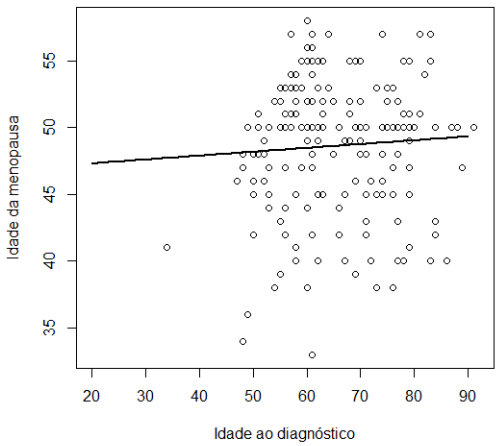



**Figura 5 - Gráfico de dispersão entre a idade de diagnóstico com a idade da menopausa**

A média da duração de amamentação varia conforme o grau de escolaridade. A média mais alta corresponde a uma escolaridade baixa. Pode-se dever ao facto

de a paciente não se encontrar a trabalhar. A média mais baixa corresponde às pacientes que têm um curso.

A amamentação, para a paciente, traz muitas vantagens. No entanto, é de importante relevância saber se o facto de a paciente amamentar altera a idade ao diagnóstico, caso seja diagnosticada a doença.

| Idade | Amamentação | |
|---|---|---|
| | Sim | Não |
| Média | 62,79 | 56,27 |
| Desvio-padrão | 13,71 | 13,82 |
| Mínimo | 27 | 28 |
| Máximo | 83 | 88 |

**Tabela 12 - Idade ao diagnóstico comparativamente à amamentação**

Como podemos observar pela Tabela 12, as médias das idades ao diagnóstico sabendo que amamentou é de 62 anos.

Comparando a idade ao diagnóstico de acordo com a duração de amamentação obtemos a tabela 13.

| Idade ao diagnóstico<br>Duração da amamentação | 15-44 | 45-54 | 55-64 | 65-74 | >74 |
|---|---|---|---|---|---|
| Média | 9,97 | 7,19 | 7,66 | 11.73 | 14,09 |
| Desvio-padrão | 10,37 | 6,94 | 8,68 | 10,04 | 13,99 |
| Mínimo | 1 | 1 | 1 | 1 | 1 |
| Máximo | 36 | 24 | 36 | 36 | 72 |

**Tabela 13- Idade ao diagnóstico comparativamente à duração da amamentação**

A duração da amamentação é distinta nas faixas etárias inferiores a 64 e superior a 64 anos.

A figura 6 mostra a correlação entre a idade ao diagnóstico e a duração da amamentação, que, como poderemos verificar, é fraca entre a idade ao

diagnóstico e a duração de amamentação em que o valor da correlação = 0.17 e o valor da prova = 0.0155

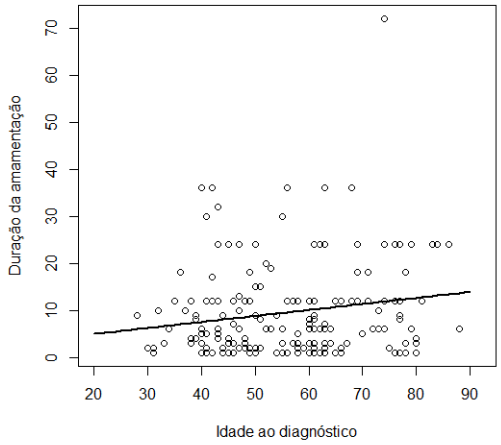



**Figura 6- Gráfico da dispersão entre a idade ao diagnóstico e a duração da amamentação**

Também fomos saber se o facto da idade da menarca da paciente altera a idade ao diagnóstico, caso seja diagnosticada a doença.

| Idade ao diagnóstico<br>Idade à menarca | 15-44 | 44-54 | 55-64 | 65-74 | >74 |
|---|---|---|---|---|---|
| Média | 13,10 | 12,90 | 13,48 | 14,37 | 14,3 |
| Desvio-padrão | 1.47 | 1.65 | 1,75 | 2,25 | 1,94 |
| Mínimo | 11 | 9 | 10 | 10 | 11 |
| Máximo | 17 | 17 | 19 | 18 | 18 |

**Tabela 14- Idade ao diagnóstico comparativamente à idade da menarca**

A Tabela 14 mostra que não há diferenças na idade ao diagnóstico comparativamente com a idade da menarca.

A figura 7, mostra a correlação entre a idade ao diagnóstico com a idade menarca

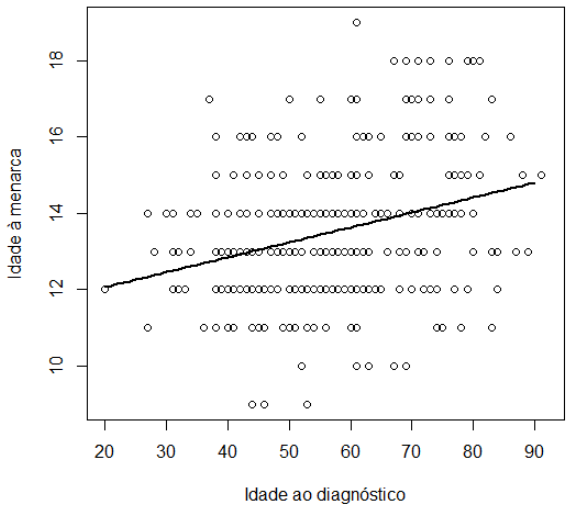



**Figura 7 - Gráfico da dispersão entre a idade do diagnóstico com a idade da menarca**

Como o valor da correlação é positivo e próximo de zero podemos concluir que a correlação é positiva fraca, em que o valor da correlação = 0.29 e o valor da prova = 9.3e-08

## *Carcinoma*

### *Tipos histológicos*

Existem vários estudos relativos ao valor prognóstico dos diferentes tipos histológicos. Na tabela 15, apresenta-se a classificação dos principais tipos histológicos de cancro da mama, segundo a Organização Mundial de Saúde.

**Tabela 15- Tipos Histológicos de cancro da mama**

| |
|---|
| - Tumores Epiteliais não invasivos |
| Carcinoma intraductal (in situ) |
| Carcionoma lobular in situ |
| - Tumores Epiteliais Invasivos |
| Carcinoma ductal invasor |
| Carcinoma ductal invasor com componente intraductal predominante |
| Carcinoma lobular invasor |
| Carcinoma mucinoso |
| Carcinoma medular |
| Carcinoma papilar |
| Carcinoma tubular |
| Carcinoma cístico adenoide |
| Carcinoma secretor (juvenil) |
| Carcinoma apócrino |
| Carcinoma com metaplasia |
| Carcinoma cribriforme invasivo |
| Carcinoma inflamatório |
| Outros |
| - Doenças do Paget do mamilo |
| - Outros tipos de tumores menos frequentes |

Os carcinomas são a maioria das neoplasias malignas da mama, sendo o ductal Invasor, o tipo mais frequente com aproximadamente 70%. Carcinoma Inflamatório apresenta uma frequência de 0.6%, sendo uma das formas mais agressivas, cuja apresentação clinica caracterizada pelo predomínio de alterações na pele da mama.

**Tabela 16 - Mama intervida**

| Mama intervida | Número (%) | Vivo (%) | Morto (%) |
|---|---|---|---|
| Direita | 250 (44,7) | 226 (40,5) | 24 (4,3) |
| Esquerda | 308 (55,2) | 272 (48,7) | 36 (6,5) |
| NA´s | 1 (0,2) | 1(0,2) | |

Como podemos observar pela tabela 16, 55.2% dos tumores foram na mama esquerda e 41,98% no quadrante superior direito.

### *Grau de diferenciação*

Bloom e Richardson [3] dividem o grau histológico em três categorias: grau I ($G_1$), (carcinoma bem diferenciado); grau II ($G_2$), (carcinoma diferenciado moderado) e grau III ($G_3$), (carcinoma mal diferenciado). O grau histológico, como fator de prognóstico, tem sido amplamente estudado e o seu impacto na sobrevivência de pacientes foi confirmado - a sobrevivência piora progressivamente à medida que aumenta o grau histológico [5].

**Tabela 17- Grau de diferenciação**

| | Quantidade (%) | Vivo (%) | Morto (%) |
|---|---|---|---|
| Grau de diferenciação | | | |
| $G_1$ | 168 (32,8) | 162 (31,6) | 6 (1,2) |
| $G_2$ | 224 (43,8) | 203 (39,6) | 21 (4,1) |
| $G_3$ | 118 (23) | 95 (18,6) | 23 (4,5) |
| $G_x$ | 2 (0,4) | 2 (0,4) | 0 (0) |
| NA´s | 47 (8,4) | 37(6,6) | 10(1,8) |
| Total | 559 | | |

Os casos em análise, apresentam predominantemente um grau histológico $G_2$, representando 43,8 % dos casos. Em menor percentagem, são os casos com grau $G_3$ (23%). Observe-se que, neste estudo, a classificação $G_x$ significa que não foi possível classificar o grau histológico do tumor e quantificam-se apenas 2 casos analisados, como podemos ver pela tabela 17.

### *HER.2*

A proteína HER (C-erbB-2, her2.neu ou oncogene neu) é um recetor transmembranar, sobre-expresso no cancro da mama, e tem atividade idêntica ao fator de crescimento epidérmico. Os tumores que sobre-expressam o HER2.neu estão habitualmente associados a um pior prognóstico e têm um valor preditivo no cancro da mama.

**Tabela 18 - Her2.neu**

| | | Número (%) | Vivo (%) | Morto (%) |
|---|---|---|---|---|
| Her.2neu | | | | |
| | Negativo | 168 (40,5) | 149 (35,9) | 19 (4,6) |
| | Positivo | 247 (59,5) | 222 (53,5) | 25 (6,0) |
| | NA´s | 144 (25,8) | 128 (22,9) | 16(2,9) |
| | Total | 559 | | |

Como se pode verificar na tabela 18, a maioria das pacientes têm uma expressão do recetor transmembranar positiva.

Comparando o recetor Her2.neu com o grau de diferenciação, para verificar se existia alguma relação (tabela 19), verificamos que, a percentagem no grau de diferenciação II quase duplica, isto é, 15,3% tem um HER.2 negativo e 28.8% tem HER.2neu positivo.

**Tabela 19 – HER2.neu com tipos histológicos**

| HER2.neu | G1 (%) | G2 (%) | G3 (%) | Gx (%) | NA's (%) |
|---|---|---|---|---|---|
| Negativo | 61 (15,5) | 60 (15,3) | 35 (8,9) | 1 (0,3) | 11 (2) |
| Positivo | 65 (16,5) | 113 (28,8) | 57 (14,5) | 1 (0,3) | 11(2) |
| NA´s | 42 (7,5) | 51 (9,1) | 26 (4,7) | 0 (0) | 25 (4,5) |

### *Recetores hormonais*

Os recetores de estrogénios (RE) e os de progesterona (RP) são fatores preditivos de resposta ao tratamento hormonal, mas também são considerados importantes indicadores de prognóstico. A tabela 20 mostra-nos o número de pacientes com os respetivos recetores hormonais.

**Tabela 20- Recetores hormonais**

| | Negativo (%) | Positivo (%) | NA's (%) |
|---|---|---|---|
| Estrogénio | 66 (13,7) | 416 (86,3) | 77 (13,8) |
| Progesterona | 114 (24,8) | 345 (75,2) | 100 (17,9) |

O valor prognóstico dos recetores hormonais tem maior incidência em pacientes que utilizaram estes mesmos recetores. Pelas tabelas 21 e 22, podemos verificar que HER2.neu. Tem maior percentagem quando o recetor de estrogénio e progesterona é positivo.

**Tabela 21- Recetores de estrogénio**

| | Recetores de estrogénio | | |
|---|---|---|---|
| HER2.neu | Negativo (%) | Positivo (%) | NA's (%) |
| Negativo | 24 (5,8) | 143 (34,8) | 1 (0,2) |
| Positivo | 37 (9,0) | 207 (50,4) | 3 (0,5) |
| NA's | 5 (0,9) | 66 (11,8) | 73 (13,1) |

**Tabela 22- Cruzamento do HER2.neu com o recetor de progesterona**

| | Recetores de progesterona | | |
|---|---|---|---|
| HER2.neu | Negativo (%) | Positivo (%) | NA´s (%) |
| Negativo | 41 (10) | 127 (30,9) | 0 (0) |
| Positivo | 66 (16,1) | 177 (43,1) | 4 (0,7) |
| NA's | 7 (1,3) | 41 (7,3) | 96 (17,2) |

### *Triplo negativo*

O triplo negativo são um subtipo de cancro da mama, o que significa que a paciente tem os recetores de estrogénio de progesterona e Her.2 negativos.

**Tabela 23 -Triplo negativo**

| | Número (%) | Vivo (%) | Morto (%) |
|---|---|---|---|
| Triplo negativo | | | |
| Sem triplo negativo | 459 (95,6) | 418 (87,1) | 41 (8,5) |
| Com triplo negativo | 21 (4,4) | 11 (2,3) | 10 (2,1) |
| NA's | 79 (14,1) | 70 (12,5) | 9 (1,6) |

O triplo negativo é um tipo de tumor mais agressivo e, como podemos ver pela Tabela 23, 95,6%, não têm triplo negativo.

A tabela 24 indica-nos o número de pacientes que tiveram recidiva de acordo com o triplo negativo.

**Tabela 24-Triplo negativo com a recidiva**

| Triplo negativo | Recidiva | |
|---|---|---|
| | Não (%) | Sim (%) |
| Sem triplo negativo | 402 (83,8) | 57 (11,9%) |
| Com triplo negativo | 11 (2,3%) | 10 (2,1) |

Podemos verificar pela Tabela 24, que, das 21 pacientes com triplo negativo, 10 tiveram recidiva.

O Ki.67 é um anticorpo monoclonal que identifica antigénios em núcleos de células em fase proliferativa. A determinação da expressão do Ki67 tem-se tornado num dos métodos preferidos para avaliação da capacidade proliferativa das células tumorais.

O índice de proliferação Ki.67 demonstrou um valor prognóstico significativo em diversos estudos do cancro da mama. A tabela 25 mostra que o número de pacientes com o valor de KI.67 alto e baixo.

**Tabela 25 - Dados descritivos do Índice de proliferação Ki.67**

| | | Número (%) | Vivo (%) | Morto (%) |
|---|---|---|---|---|
| Ki.67 | | | | |
| | Alto | 178 (56,7) | 158 (50,3) | 20 (6,4) |
| | Baixo | 136 (43,3) | 129 (41,1) | 7 (2,2) |
| | NA's | 245 (43,8) | 212 (37,9) | 33 (5,9) |

Podemos constatar que a percentagem de paciente com Ki.67 é superior quando o índice de proliferação é alto. Comparando por grau histológico, no $G_1$ a quantidade é maior quando o índice de proliferação é menor, não acontecendo com o $G_2$ nem com o G3. A tabela 26 mostra-nos o cruzamento do Ki.67 com o grau histológico.

**Tabela 26 - Cruzamento Ki.67 com grau histológico**

| | $G_1$ (%) | $G_2$ (%) | $G_3$ (%) | $G_x$ (%) | NA's (%) |
|---|---|---|---|---|---|
| Ki.67 | | | | | |
| Alto | 24 (8) | 82 (27,2) | 62 (20,6) | 1 (0,3) | 9 (1,6) |
| Baixo | 73 (24,3) | 50 (16,6) | 8 (2,7) | 1 (0,3) | 3 (0,7) |
| NA's | 71 (12,7) | 92 (16,5) | 48 (8,6) | 0 (0) | 34 (6,1) |

No que respeita à recidiva, quer a paciente tenha um índice de proliferação alto ou baixo, este não parece ter influência na recidiva. A tabela 27 mostra o

número de pacientes que tiveram recidiva com o cruzamento do Índice de proliferação alto ou baixo.

**Tabela 27 - Cruzamento Ki.67 com recidiva**

| | Sem recidiva (%) | Com recidiva (%) |
|---|---|---|
| Ki.67 | | |
| Alto | 151 (48,1) | 27 (8,6) |
| Baixo | 126 (40,1) | 10 (3,2) |
| NA's | 201 (36) | 44 (7,9) |

*Cirurgia*

Do total de pacientes que se encontravam em fase de tratamento, fomos verificar se a paciente fez ou não cirurgia e qual o tipo de cirurgia feita.

**Tabela 28- Cirurgia**

| | Número | Vivo (%) | Morto (%) |
|---|---|---|---|
| Cirurgia | | | |
| Não foi submetido | 14 (2,5) | 9 (1,6) | 5 (0,9) |
| Foi submetido | 545 (97,5) | 490 (87,7) | 55 (9,8) |
| Tipo cirurgia | | | |
| - Conservadora | 224 (40,1) | 213 (38,1) | 11 (2) |
| - Mastectomia | 321 (57,4) | 277 (49,6) | 44 (7,9) |

A tabela 28, mostra-nos que, 97% destas foram sujeitas a cirurgia. A cirurgia mais efetuada foi a mastectomia, com 57,4%. É de salientar que, das pacientes que foram sujeitas à cirurgia, quase 98% se encontravam vivas no término deste estudo.

Saber se a paciente teve recidiva é também um fator importante de estudo. A Tabela 29, apresenta o número de casos de recidiva, bem como a distribuição pelos tipos de recidiva.

**Tabela 29 - Tipo de recidiva**

| | Número (%) | Vivo (%) | Morto (%) |
|---|---|---|---|
| Recidiva | | | |
| Sim | 81 (14,5) | 29 (5,2) | 52 (9,3) |
| Não | 478 (85,5) | 470 (84) | 8 (1,4) |
| Tipo de recidiva | | | |
| 0 -Sem recidiva | 478 (85,5) | 470 (84,1) | 8 (1,4) |
| 1 –Recidiva loco-regional | 11 (2,0) | 7(1,3) | 4 (0,7) |
| 2 –Recidiva à distância | 64 (11,4) | 21(3,8) | 43 (7,7) |
| 3 –Recidiva mista | 6 (1,1) | 1 (0,2) | 5 (0,9) |

Como podemos observar na tabela 29, a sua grande maioria não teve recidiva. E as que tiveram recidiva, a sua maioria foi recidiva à distância.

É de salientar que a recidiva mista foi mais mortífera que comparativamente às outras recidivas.

### *Historial Familiar*

O cancro da mama é o mais comum em Portugal. No entanto, o facto de ter algum familiar com a doença não significa necessariamente que tenha maior probabilidade de a desenvolver. A maior parte dos cancros da mama não se devem a fatores genéticos e não afeta o risco de outros familiares.

**Tabela 30 - Historial familiar**

| | Número (%) | Vivo (%) | Morto (%) |
|---|---|---|---|
| Historial familiar | | | |
| Não | 50 (42,4) | 41 (34,7) | 9 (7,6) |
| Sim | 68 (57,6) | 63 (53,4) | 5 (4,2) |
| NA´S | 441 (78,9) | | |
| Grau parentesco | | | |
| 1 (pai/ filho e mãe) | 18 (26,9) | 16 (23,9) | 2 (3,0) |

| | | | |
|---|---|---|---|
| 2 (irmãos e avós) | 24 (35,8) | 23 (34,3) | 1 (1,5) |
| 3 (tios, sobrinhos e bisavós) | 16 (23,9) | 15 (22,4) | 1 (1,5) |
| 4 (primos e trisavós) | 9 (13,4) | 8 (11,9) | 1 (1,5) |
| NA´s | 492 (88,0) | 437 (78,2) | 55 (9,8) |

Como poderemos verificar pela Tabela 30, apesar de quase 60% das pacientes terem familiar com cancro da mama, este poderá não estar relacionado com o grau de parentesco. Também temos de referir que destes 60%, 53,4% ainda se encontrava vivo ao término deste estudo.

### *Mamografia / Ecografia*

O principal exame de rastreamento do cancro da mama é a mamografia. A principal função deste exame é diagnosticar lesões precoces, não palpáveis em estágio subclínico. Este exame é bastante importante pois aumenta a probabilidade de sucesso no tratamento caso este seja descoberto. A ecografia/ultrassonografia mamária é um exame complementar da mamografia.

Aproximadamente 99%das pacientes efetuaram uma mamografia/ ecografia anterior à consulta de diagnóstico, não sabendo, contudo, se esta foi por indicação médica por desconfiança ou se foi apenas por rotina de prevenção.

### *Tratamento primário*

Um tratamento é chamado neoadjuvante quando é administrado antes do tratamento definitivo, em geral cirúrgico ou, mais raramente radioterápico. Existem quatro modalidades de tratamento: quimioterapia, hormonoterapia, radioterapia e terapia-alvo.

No caso do cancro da mama, a grande vantagem do tratamento neoadjuvante é tentar reduzir o tamanho do tumor para tentar evitar a mastectomia (retirada cirúrgica completa da mama).

**Tabela 31 - Dados relativos ao tratamento primário**

| | Número (%) | Vivo (%) | Morto (%) |
|---|---|---|---|
| Não fez tratamento neoadjuvante | 492 (88) | 452 (80,9) | 40 (7,2) |
| Fez tratamento neoadjuvante | 67 (12) | 47 (8,4) | 20 (3,6) |

Como podemos constatar pela tabela 31, a maioria não fez tratamento primário.

### *Presença de gânglios*

A presença de gânglios também é importante para a análise médica.

**Tabela 32 - Dados relativos à presença de gânglios**

| | Número (%) | Vivo (%) | Morto (%) |
|---|---|---|---|
| Não apresentam linfonodos | 312 (55,8) | 261 (46,7) | 51 (9,1) |
| Apresentam linfonodos | 247 (44,2) | 238 (42,6) | 9 (1,6) |
| NA´s | 0 (0) | 0 (0) | 0 (0) |

Como podemos verificar pela tabela 32, 247 pacientes apresentavam linfonodos, que são pequenas estruturas que funcionam como filtros para as substâncias nocivas.

### *Tratamento hormonal de substituição (THS)*

Na Europa, os medicamentos convencionais usados na THS (contendo estrogénios isolados ou estrogénios em associação com progesterona) estão autorizados para alívio dos sintomas da menopausa, incluindo afrontamentos, secura vaginal e suores noturnos. Alguns destes medicamentos estão também autorizados para uso na prevenção da osteoporose, sendo para esse efeito utilizados em tratamentos prolongados.

No nosso estudo, 69 pacientes não utilizaram o tratamento hormonal de substituição enquanto 26 utilizaram este mesmo tratamento.

# Capítulo 5- Estudo de sobrevivência

## 5.1. Análise de Kaplan-Meier

Dada a existência de censura, os dados de sobrevivência são resumidos de forma conveniente através de estimativas da função de sobrevivência e da função de risco (ou função $hazard$).

No estudo que se segue, apresentaremos um estimador de Kaplan-Meier para a função de sobrevivência aplicada ao cancro da mama. Trata-se de métodos não-paramétricos, porque a estimação é feita sem que se faça nenhuma suposição sobre a distribuição da probabilidade do tempo de sobrevivência. Iremos, também, utilizar alguns métodos não paramétricos para a comparação de curvas de sobrevivência, nomeadamente o teste $log - rank$.

Em ambiente R tal teste pode ser efetuado recorrendo à função $survfit$ que oferece várias opções e funcionalidades.

### Resultados da data do diagnóstico até à data da recidiva

Com o interesse de estudar os fatores de risco para a recidiva de cancro da mama estudou-se também o tempo desde a data do diagnóstico utilizando, por isso, como variável de interesse o tempo desde a data do diagnóstico até à data da recidiva. A seguir está representada a tabela 33 que nos indica o número de pacientes que tiveram recidiva e o tipo de recidiva.

**Tabela 33- Número de pacientes com recidivas e tipo de recidiva**

| | | Quantidade | % |
|---|---|---|---|
| Recidiva | Não | 478 | 85,5 |
| | Sim | 81 | 14,5 |
| Tipo de recidiva | Sem recidiva | 478 | 85,5 |
| | Recidiva loco-regional | 11 | 2 |
| | Recidiva à distância | 64 | 11,4 |
| | Recidiva mista | 6 | 1,1 |

Neste mesmo estudo, foram excluídas 12 pacientes por terem tido recidiva antes do diagnóstico.

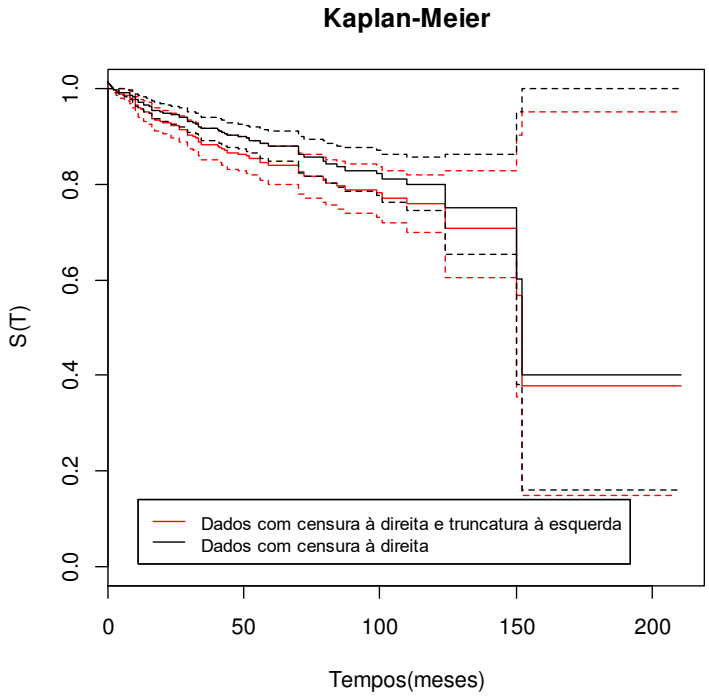



**Figura 8- Curva de Kaplan-Meier para as pacientes com recidiva do Hospital de Braga**

A figura 8, apresenta as curvas de sobrevivência de Kaplan-Meier para todas as pacientes. Considerando a linha a tracejado como a curva de sobrevivência dos dados censurados à direita e a linha sólida como a curva de sobrevivência dos dados censurados à direita e truncados à esquerda, podemos verificar que a estimativa de Kaplan-Meier de tempo até à recidiva para 120 meses é muito próxima de 80%.

Além disso, parece que, para todo o seguimento observado a probabilidade de sobrevivência é superior a 40%.

A tabela 34, mostra os valores de prova obtidos nos testes de$log-rank$ e de Wilcoxon, sob a hipótese nula de que as curvas de sobrevivência nas diferentes categorias são iguais.

Apenas as variáveis cujas curvas de sobrevivência entre as suas categorias são significativamente diferentes são apresentadas, as restantes não rejeitaram a hipótese nula, para um nível de significância de 5%.

**Tabela 34- Testes log-rank e de Wilcoxon utilizados para testar a igualdade das curvas de sobrevivência.**

| | Testes (valor prova) | |
|---|---|---|
| | $log-rank$ | Wilcoxon |
| Estadio do tumor | <0.001 | <0.001 |
| Tratamento neoadjuvante | <0.001 | <0.001 |
| Tipo de cirurgia | <0.001 | <0.001 |
| Pesq. Gg Sentinela | 0,004 | 0.004 |
| Esvaziamento axilar | 0.021 | 0,01 |
| Ki.67 | 0.002 | 0.002 |
| Invasão vascular venosa | <0.001 | <0.001 |
| Invasão vascular linfática | 0,003 | 0.003 |
| Recetores de estrogénio | 0.007 | 0.005 |
| Recetores de progesterona | 0.006 | 0.004 |
| Triplo negativo | <0.001 | <0.001 |
| Grau do tumor | 0.001 | 0.0003 |
| Tamanho do tumor | <0.001 | <0.001 |
| Gânglios regionais | <0.001 | <0.001 |
| Idade ao diagnóstico | 0.002 | 0.001 |
| Hormonoterapia | 0,002 | 0.001 |
| Gânglios Regionais | <0.001 | <0.001 |

Comparativamente ao evento morte, existem algumas diferenças relativas ao evento, recidiva, nomeadamente:

- A variável bilateral tem influência na probabilidade de morrer de cancro da mama, mas não tem influência na probabilidade de recidiva do tumor;
- As variáveis esvaziamento axilar e Ki.67 têm efeito significativo no evento recidiva mas não para o evento morte [4]

A partir dos resultados da tabela 34 e observando a figura 9, podemos aceitar que o estadio do tumor é um fator de risco significativo para a sobrevivência dessas pacientes para o risco de recidiva. O resultado do teste$log-rank$confirma que não há diferenças significativas entre a taxa de sobrevivência entre as três

primeiras categorias do estadio do tumor (0, I, II) relativamente às outras duas categorias (III, IV), como veremos mais a frente.

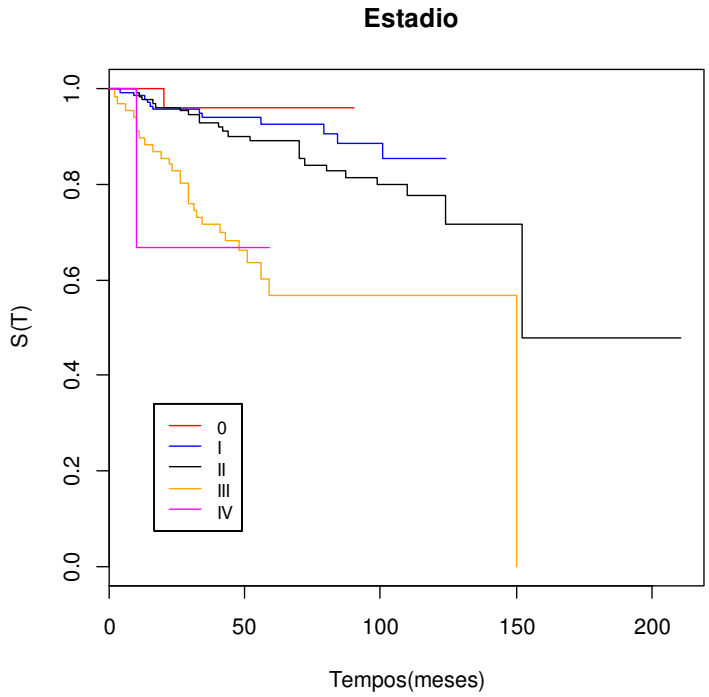



**Figura 9- Curva de Kaplan-Meier para a variável estadio**

O tratamento neoadjuvante foi inicialmente utilizado no tratamento sistémico do cancro da mama localmente avançado e inoperável.

As pacientes que foram sujeitas a este tratamento têm uma menor probabilidade de recidiva comparativamente às que não fizeram este tratamento.

Relativamente ao tipo de cirurgia efetuada podemos constatar que as pacientes que efetuaram uma mastectomia têm uma maior probabilidade de recidiva do que aquelas que efetuaram uma cirurgia conservadora.

O risco aumentado poderá estar relacionado como facto de a biópsia permitir um melhor diagnóstico e consequentemente melhor tratamento.

As pacientes que fizeram um esvaziamento axilar têm um maior risco de recidiva do que aquelas que não fizeram o esvaziamento axilar. Também podemos observar que pacientes em que o Ki.67 é alto têm uma menor probabilidade de ter recidiva.

Quando as imagens, sejam elas linfáticas ou venosas, são visíveis, a paciente tem um maior risco de ter recidiva.

Neste estudo podemos verificar que, para uma expressão de estrogénio e progesterona, positiva o risco de ter recidiva é menor.

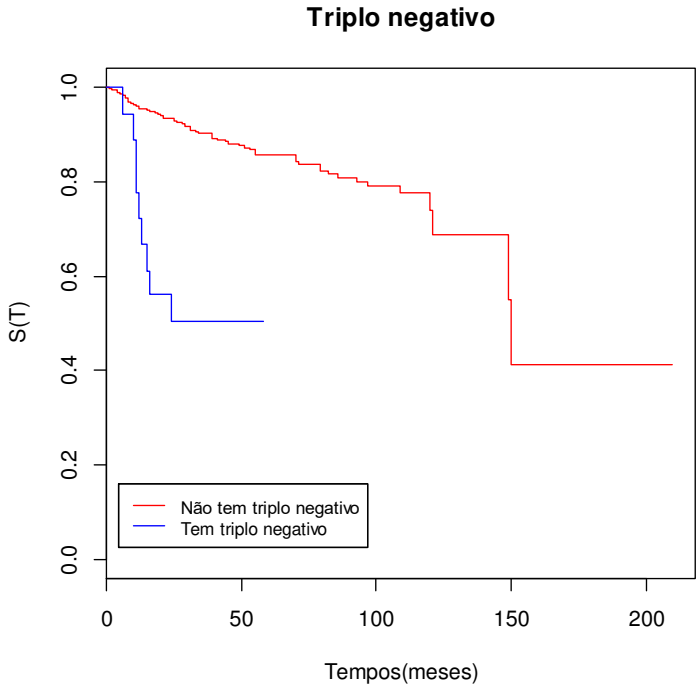



**Figura 10 - Curva de Kaplan-Meier para a variável triplo negativo**

Pela figura 10, podemos verificar que as pacientes que tiveram triplo negativo têm uma maior probabilidade de ter recidiva.

Na figura 11, podemos verificar que não existe uma diferença entre as três categorias do grau ($G_1$, $G_2$, $G_x$), como veremos mais à frente. Tumor com categoria $G_3$ tem uma menor probabilidade de não ter recidiva.

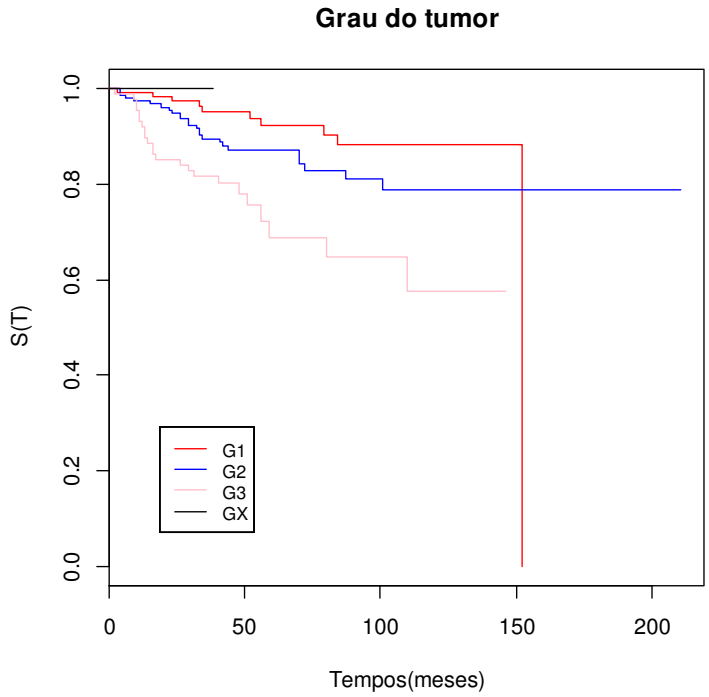



**Figura 11 - Curva de Kaplan-Meier para a variável grau do tumor**

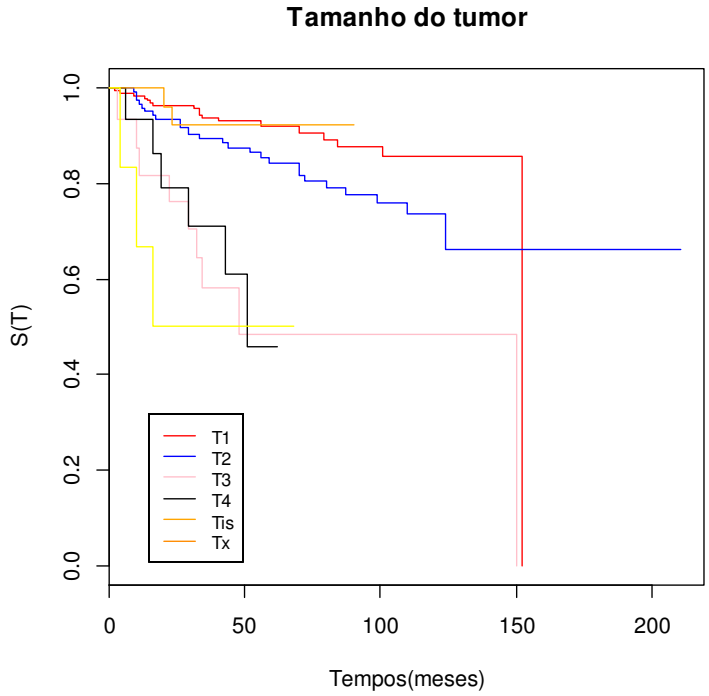



**Figura 12- Curva de Kaplan-Meier para a variável tamanho do tumor**

A figura 12 mostra-nos que quanto maior for o tumor maior será a probabilidade da paciente ter recidiva.

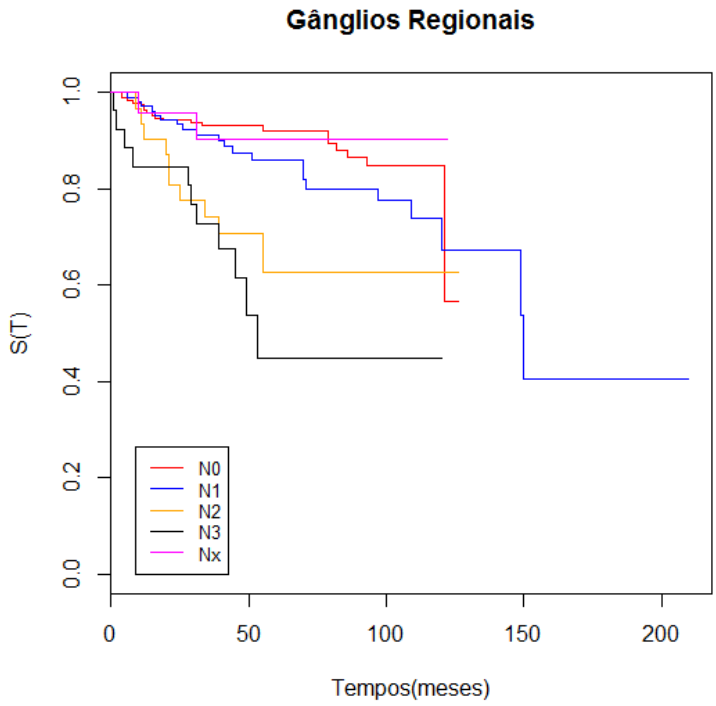



**Figura 13- Curva de Kaplan-Meier para a variável gânglios regionais**

Após análise da figura 13, podemos verificar que pacientes com tumores nas categorias $N_x$, $N_0$, $N_1$ ou $N_2$ têm uma maior probabilidade de não terem recidiva.

Esta análise será feita posteriormente na recodificação desta mesma variável.

A variável idade, categorizada, não foi estatisticamente significativa em termos de efeito de sobrevivência de recidiva, figura 14.

No entanto, quando a classificamos em duas categorias, pacientes com mais de 44 anos e pacientes com igual ou mais de 44 anos (como referenciado no RORENO [18]), verificou-se que as pacientes com menos de 44 anos tinham uma maior probabilidade de ter recidiva, como veremos mais à frente no gráfico 18.

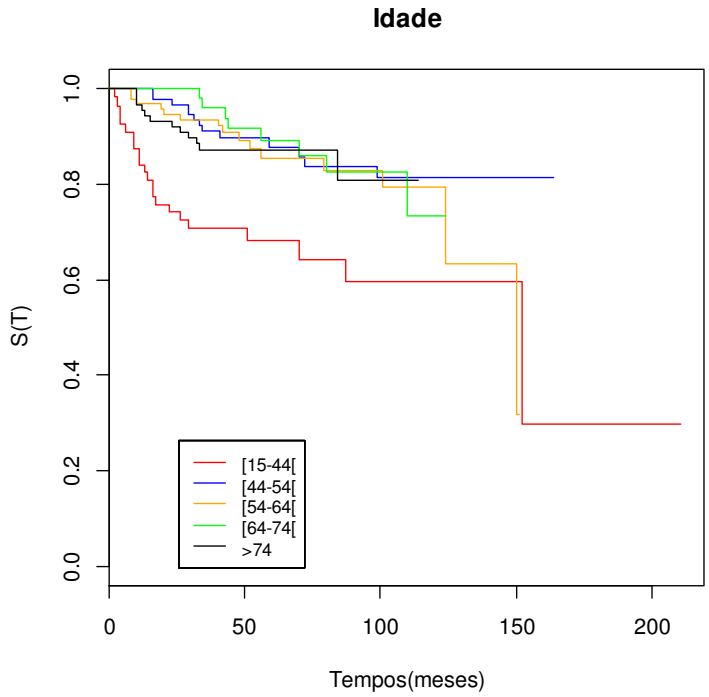



**Figura 14 -Curva de Kaplan-Meier para a variável Idade**

As pacientes que foram submetidas à hormonoterapia, têm uma maior probabilidade de não ter recidiva comparativamente às que não foram submetidas à hormonoterapia.

Tendo em conta os resultados decidiu-se recodificar algumas das variáveis com mais do que duas categorias, obtendo-se o resultado da tabela 35.

**Tabela 35- Testes de log-rank e de Wilcoxon para testar a igualdade das curvas de sobrevivência entre os grupos das variáveis categorizadas.**

| | Testes (valor-prova) | |
|---|---|---|
| | $\log - rank$ | Wilcoxon |
| Grau | <0.001 | <0.001 |
| Gânglios Regionais | <0.001 | <0.001 |
| Estadio | <0.001 | <0.001 |
| Idade | <0.001 | <0.001 |
| Tamanho do tumor | <0.001 | <0.001 |

Pelo gráfico 8, podemos observar que existe uma diferença entre a categoria $G_1$, $G_2$ e $G_x$ e a categoria $G_3$ com um valor de prova <0.001.

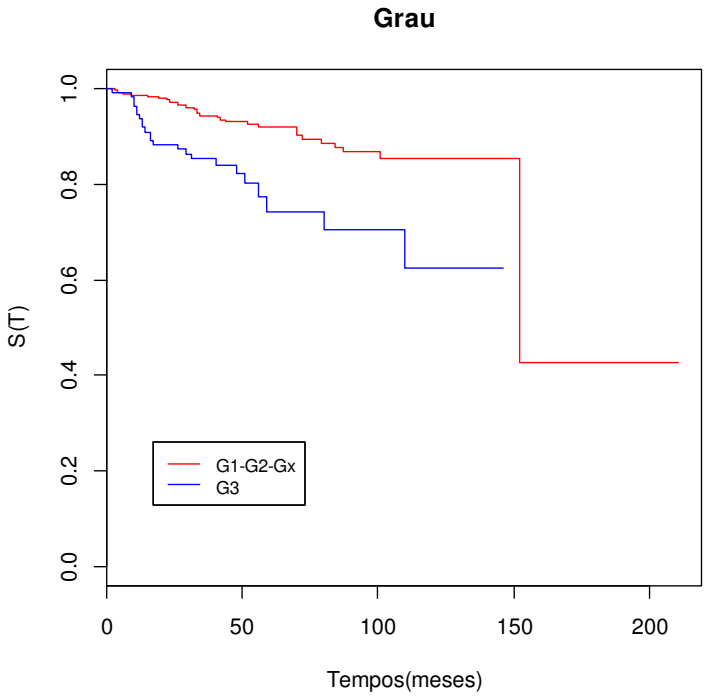



**Figura 15 - Curva de Kaplan-Meier para a variável categorizada grau**

As pacientes com o grau $G_3$, têm uma menor probabilidade de não ter recidiva.

De modo análogo, podemos verificar que, as pacientes, com $N_0$, $N_1$, $N_2$ e $N_x$, têm uma menor probabilidade de ter recidiva comparativamente as pacientes com $N_3$.

Após o teste log-rank a estas duas categorias verificamos que não existem diferenças significativas entre as curvas, com valor de prova <0.001.

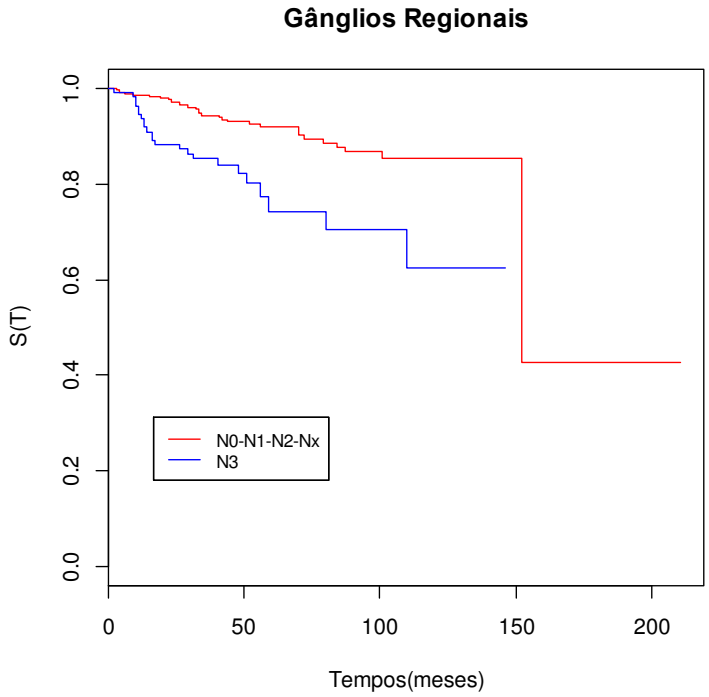



**Figura 16 - Curva de Kaplan-Meier para a variável categorizada gânglios regionais**

A variável estadio foi dividida em duas categorias, uma com o estadio 0, I e II, e a outra variável com o estadio III e IV.A Figura 16 apresenta as curvas de Kaplan-Meier para as novas categorias da variável.

Após o teste $log - rank$ podemos verificar que existem diferenças significativas nestas mesmas categorias, uma com o valor de prova <0.001.

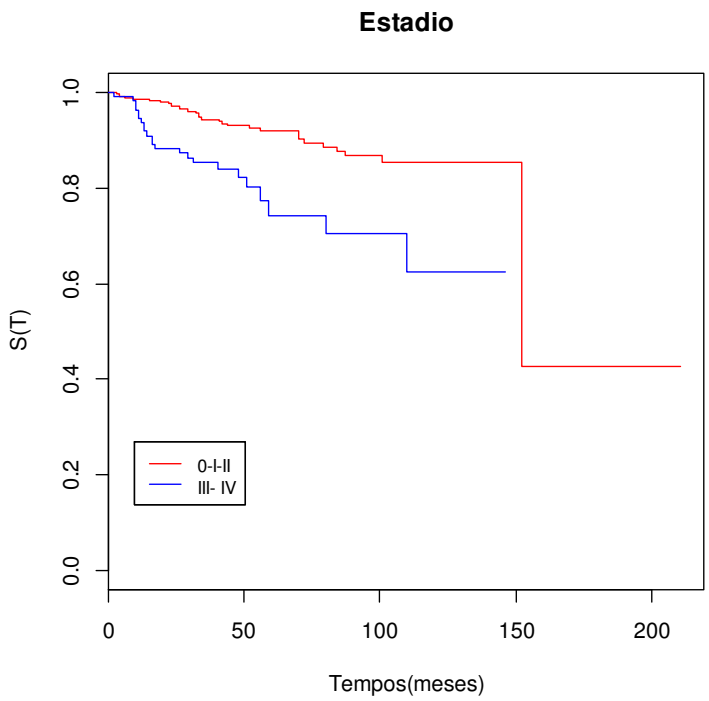



**Figura 17- Curva de Kaplan-Meier para a variável categorizada estadio**

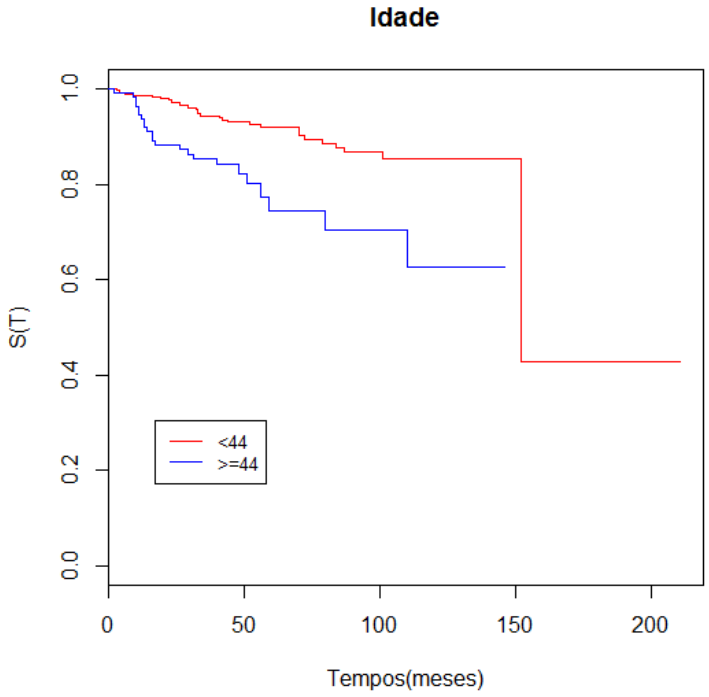



**Figura 18- Curva de Kaplan-Meier para a variável categorizada idade**

A categorização desta variável já foi referida anteriormente, pelo que o teste $log-rank$ obtém um valor de prova <0,001, respetivamente. Claro que pelo gráfico podemos observar que existe diferenças significativas.

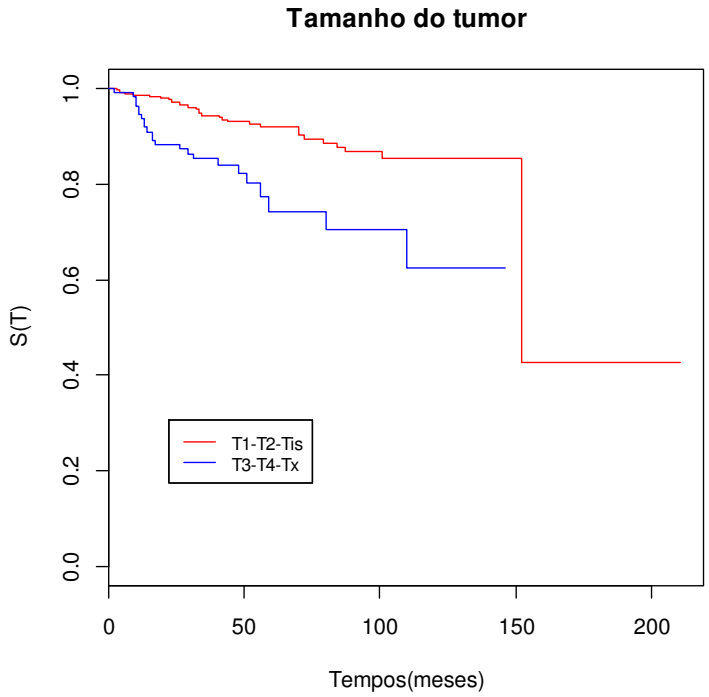



**Figura 19- Curva de Kaplan-Meier para a variável categorizada, tamanho do tumor**

A figura 19 sugere que pacientes com um tamanho do tumor $T_1$, $T_2$, $T_{is}$ têm menor probabilidade de ter recidiva, com valor prova <0.001.

## Resultados da data do tratamento até à recidiva

Foram retiradas 14 pacientes da nossa base de dados, devido a estas terem tido recidiva antes do início do nosso estudo ou terem tido recidiva anteriormente à data de tratamento.

Com o interesse de estudar os fatores de risco para a recidiva de cancro da mama estudou-se também o tempo desde a data do início do tratamento, utilizando, por isso, como variável de interesse o tempo desde o início do tratamento até à recidiva por cancro da mama

**Tabela 36- Testes de $log-rank$ e de Wilcoxon utilizados para testar a igualdade das curvas de sobrevivência**

| | Testes (valor-prova) | |
|---|---|---|
| | $log-rank$ | Wilcoxon |
| Estadio do Tumor | <0.001 | <0.001 |
| Tratamento neoadjuvante | <0.001 | <0.001 |
| Tipo de cirurgia | <0.001 | <0.001 |
| Pesquisa. Gânglios Sentinela | 0.003 | 0.003 |
| Esvaziamento axilar | 0.024 | 0.019 |
| Ki.67 | 0.002 | 0.001 |
| Invasão vascular venosa | <0.001 | <0.001 |
| Invasão vascular linfática | 0.001 | 0.001 |
| Recetores de estrogénio | 0.007 | 0.005 |
| Recetores de progesterona | 0.005 | 0.003 |
| Triplo negativo | <0.001 | <0.001 |
| Grau do tumor | 0.000 | 0.006 |
| Tamanho do tumor primário | <0.001 | <0.001 |
| Gânglios regionais | <0.001 | <0.001 |
| Idade ao diagnóstico | 0.001 | 0.000 |
| Hormonoterapia | 0.001 | 0.001 |

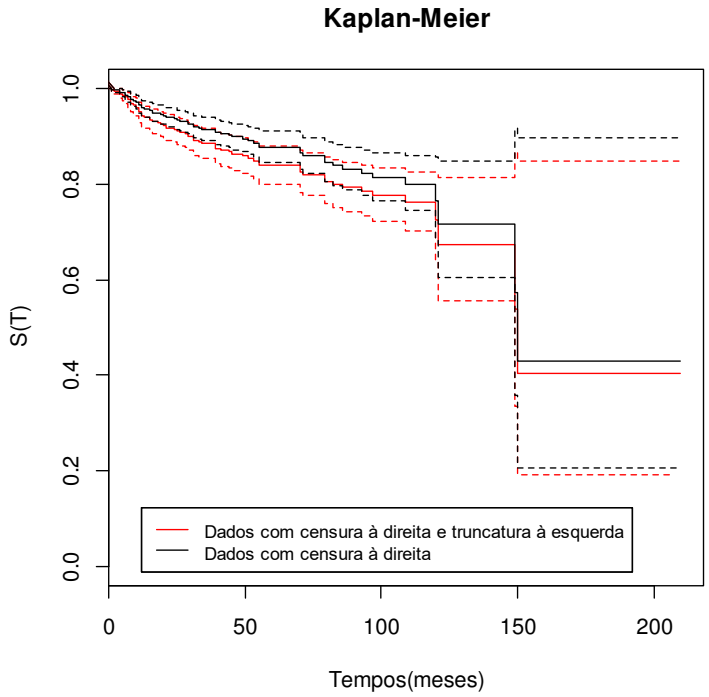



**Figura 20- Curva de Kaplan-Meier para os dados**

A figura 20 apresenta as curvas de sobrevivência para todas as pacientes desde o tempo de tratamento até à morte.

A linha a vermelho corresponde aos dados apenas censurados à direita e a linha a preto corresponde aos dados censurados à direita e truncados à esquerda. Como podemos verificar pelo gráfico a estimativa de Kaplan-Meier para a sobrevivência a recidiva a 120 meses destas pacientes é muito próxima de 40%.

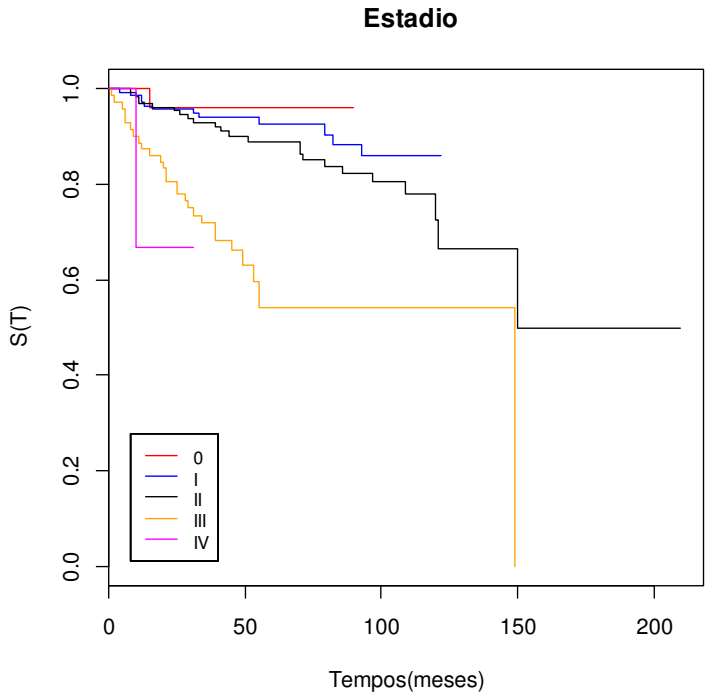



**Figura 21 - Curva de Kaplan-Meier para a variável estadio**

A partir dos dados da tabela 36, e da figura 21, podemos concluir que o estadio do tumor é um fator de risco significativo para a sobrevivência das pacientes desde o início do tratamento até à recidiva.

Após o teste $log - rank$ confirma que há diferenças significativas na taxa de sobrevivência entre as três primeiras categorias (0-I-II) assim como para as outras duas categorias (III-IV) com um valor de prova <0.001.

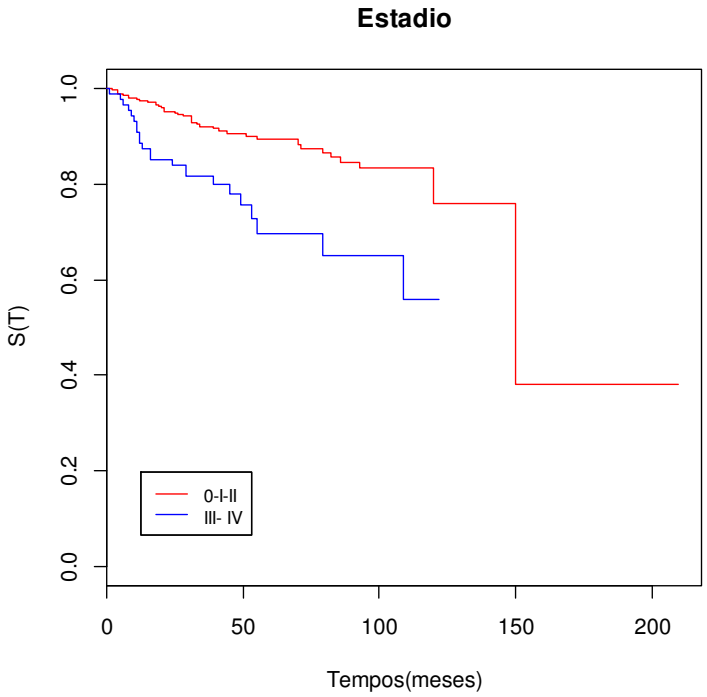



**Figura 22 - Curva de Kaplan-Meier para a variável recodificada estadio**

Como podemos observar pela figura 22, as pacientes que efetuaram tratamento primário, têm uma maior probabilidade de morte comparativamente às que não fizeram tratamento.

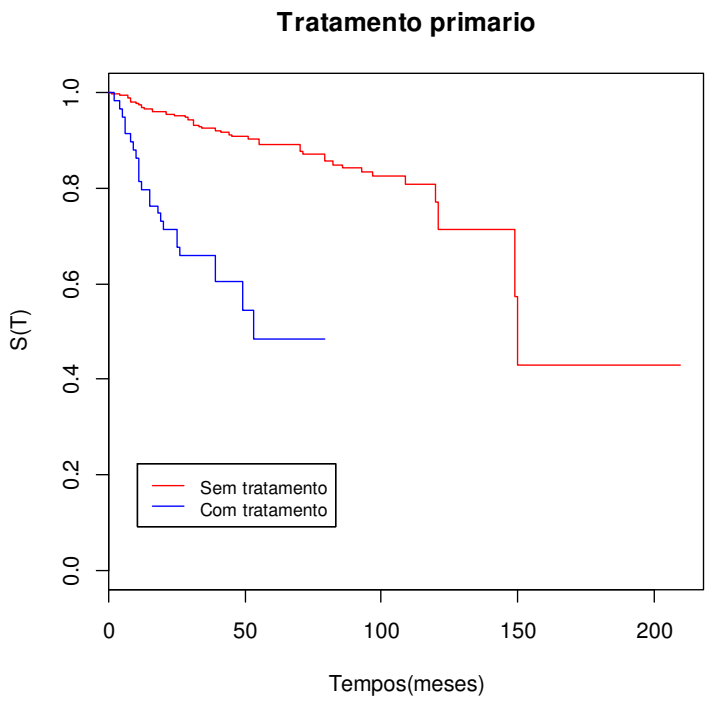



**Figura 23 - Curva de Kaplan-Meier para a variável tratamento primário**

Relativamente ao tipo de cirurgia efetuada podemos constatar que as pacientes que efetuaram mastectomia têm uma maior probabilidade de recidiva do que aquelas que efetuaram uma cirurgia conservadora.

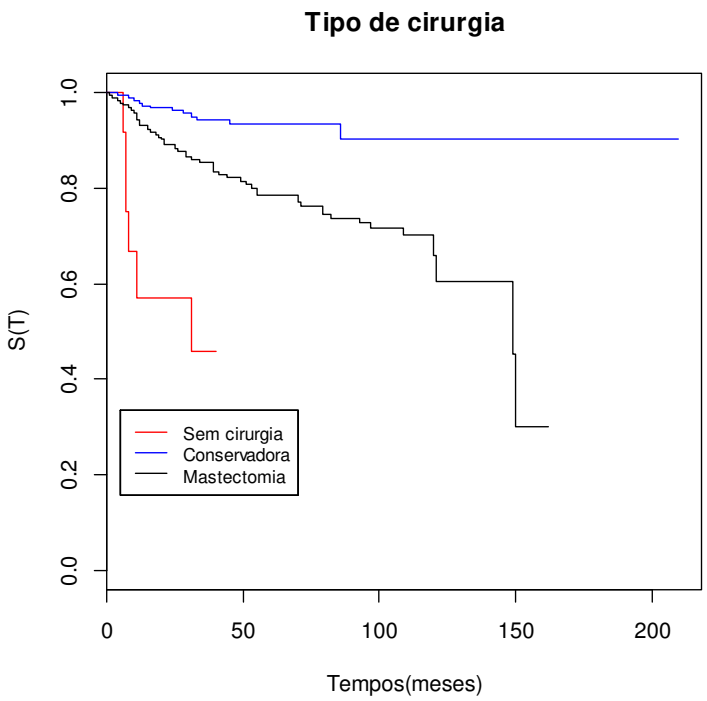



**Figura 24- Curva de Kaplan-Meier para a variável tipo de cirurgia**

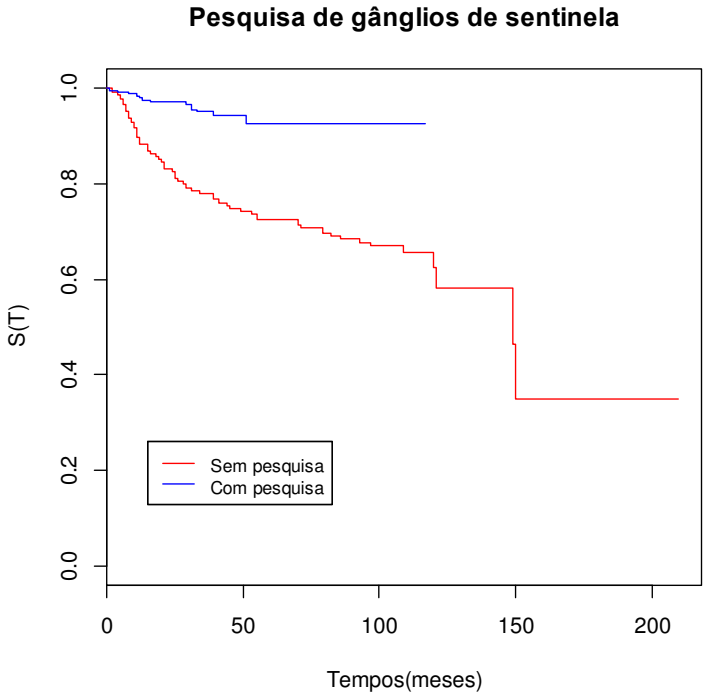



**Figura 25- Curva de Kaplan-Meier para a variável pesquisa de gânglios de sentinela**

Relativamente à biopsia e pela figura 25, podemos verificar que quem não fez essa biópsia tem um maior risco de ter recidiva.

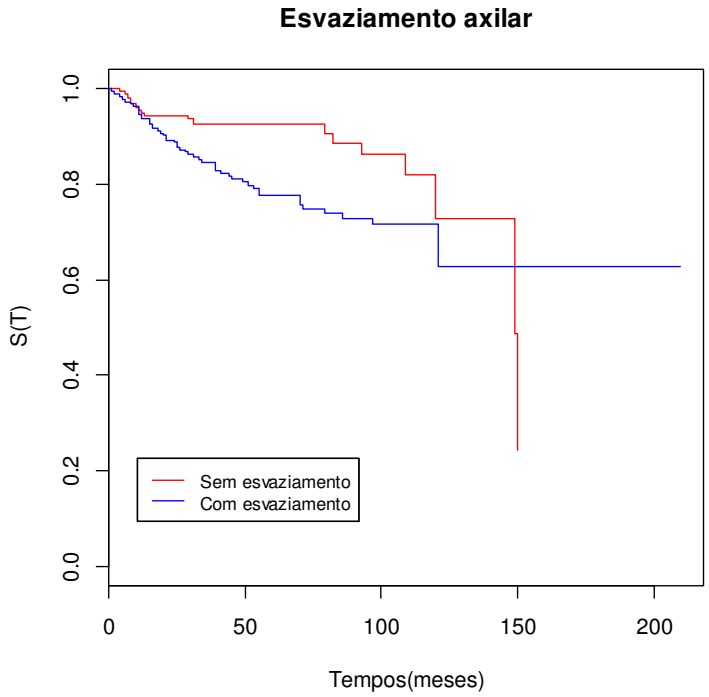



**Figura 26 - Curva de Kaplan-Meier para a variável esvaziamento axilar**

As pacientes a que foram sujeitas ao esvaziamento axilar têm um maior risco de recidiva do que aquelas que não foram sujeitas, como poderemos verificar pela figura 26.

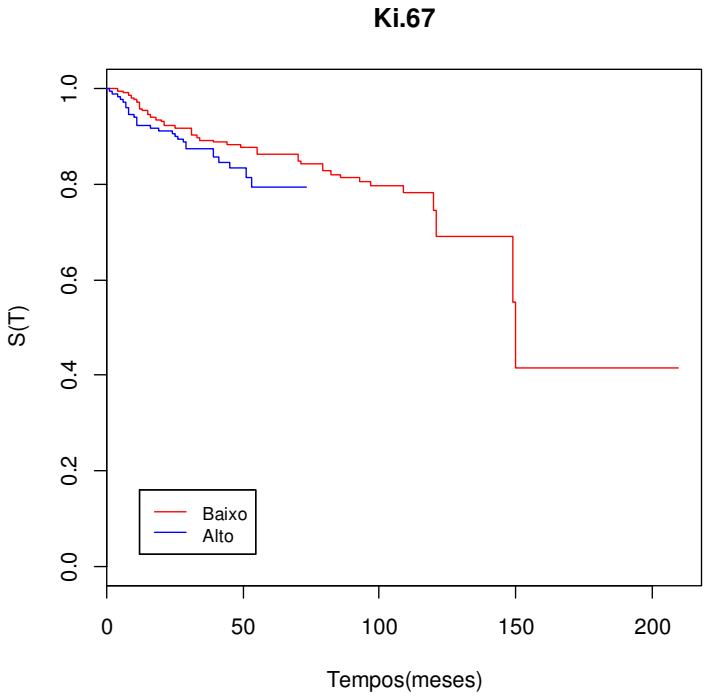



**Figura 27- Curva de Kaplan-Meier para a variável KI.67**

Na figura 27, podemos observar que as pacientes em que, o Ki.67 é alto, a probabilidade de ter recidiva é menor.

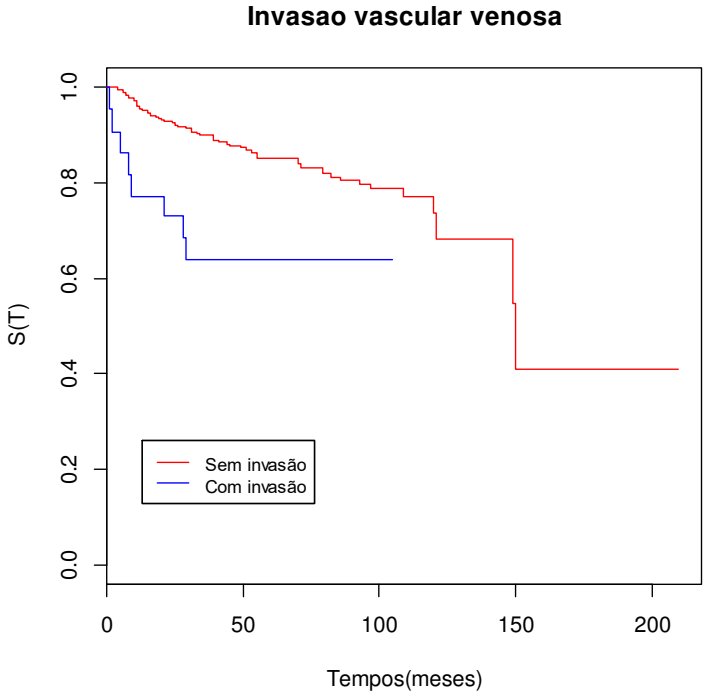



**Figura 28- Curva de Kaplan-Meier para a variável invasão vascular venosa**

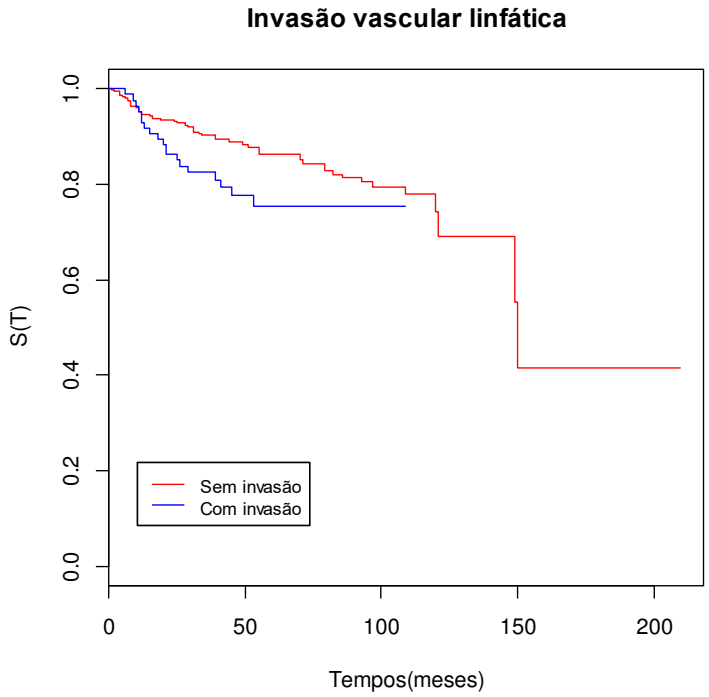



**Figura 29- Curva de Kaplan-Meier para a variável invasão vascular linfática**

Pelas figuras 28 e 29 podemos verificar que quando as imagens, sejam elas linfáticas ou venosas, sejam visíveis, a paciente tem um maior risco de ter recidiva.

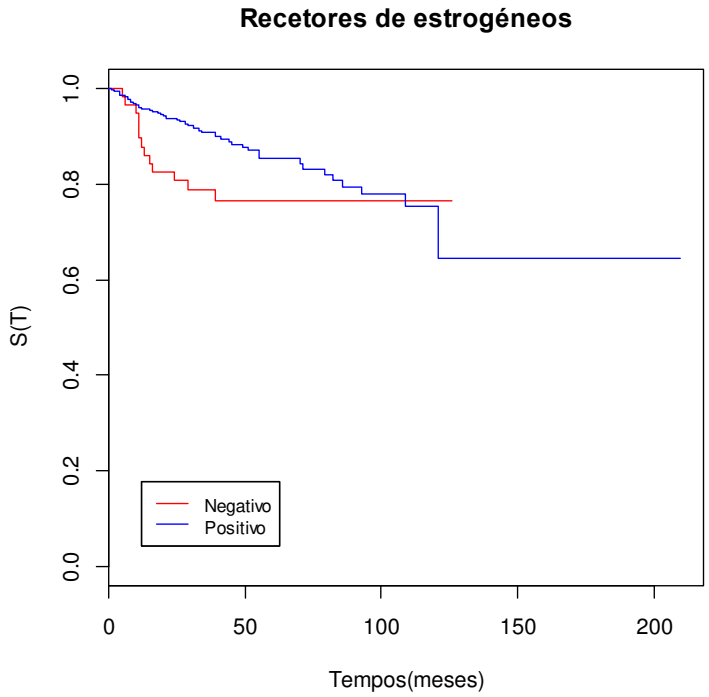



**Figura 30- Curva de Kaplan-Meier para a variável recetores de expressão de estrogénio**

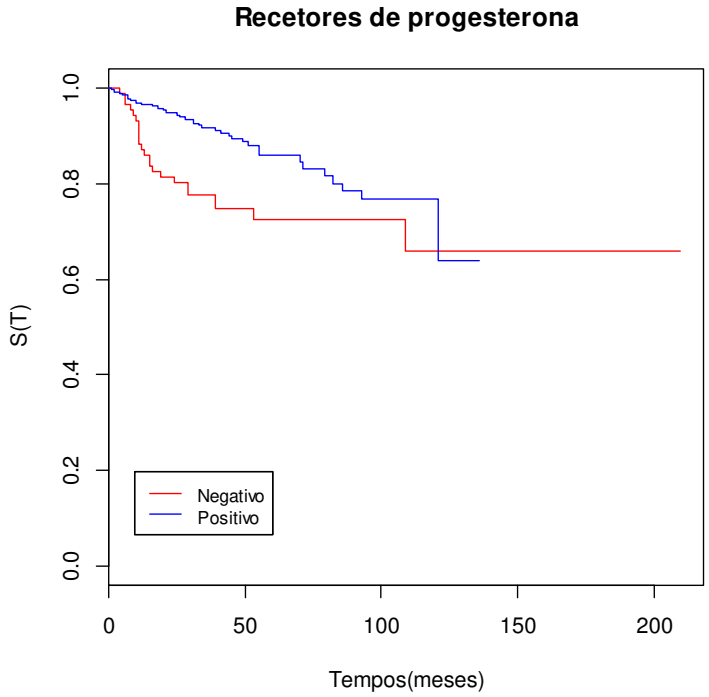



**Figura 31 -Curva de Kaplan-Meier para a variável recetora de expressão de progesterona**

Similarmente aos anteriores, as figuras 30 e 31, mostram que em ambos os casos, expressão de estrogénio e expressão de progesterona, sejam positivas, o risco de ter recidiva é menor.

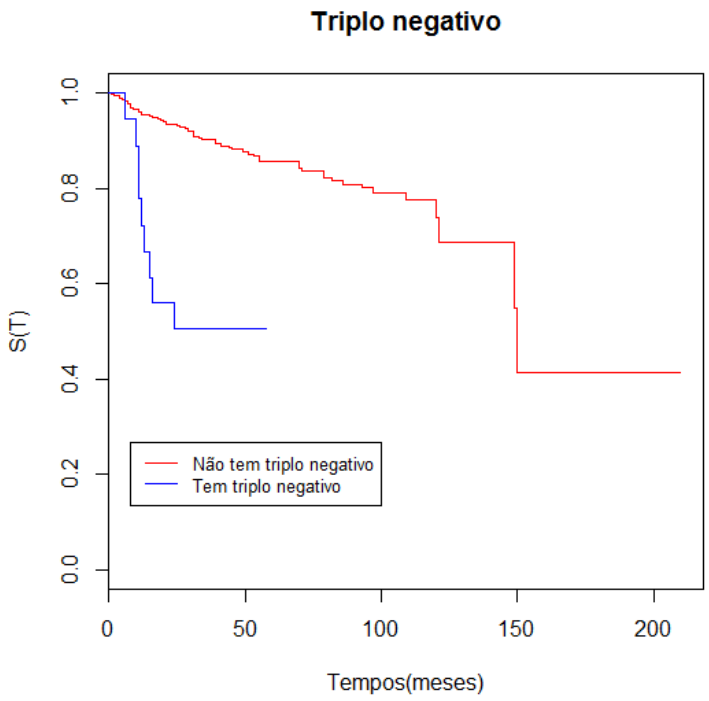



**Figura 32 - Curva de Kaplan-Meier para o variável triplo negativo**

A figura 32, podemos verificar que as pacientes que tiveram triplo negativo têm uma maior probabilidade de ter recidiva.

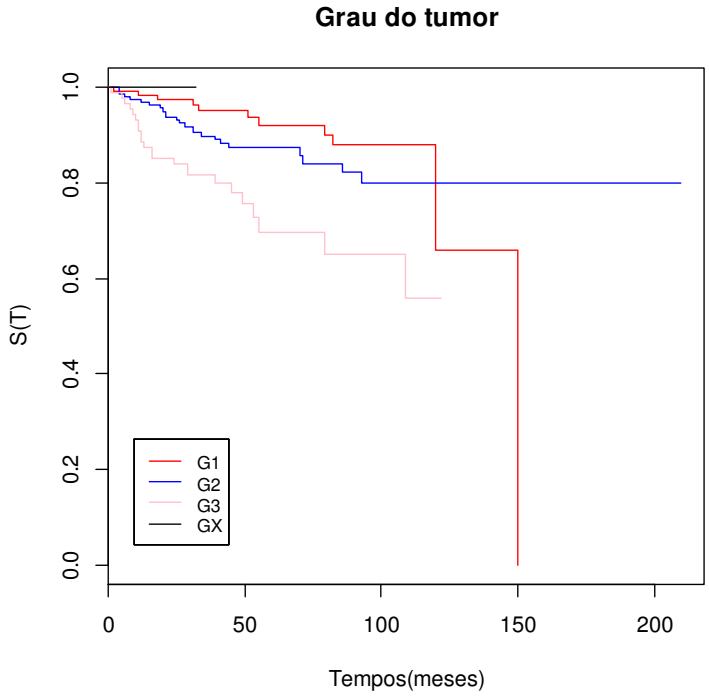



**Figura 33 - Curva de Kaplan-Meier para a variável grau do tumor**

A figura 33, mostra que não existe diferença nas três categorias do grau ($G_1$, $G_2$, GX). Também podemos observar que quem tem o grau com a categoria $G_3$ tem uma menor probabilidade de não ter recidiva. Após o teste $log - rank$ podemos verificar que existem diferenças significativas entre as categorias ($G_1$, $G_2$, Gx) com $G_3$, com um valor de prova <0,001.

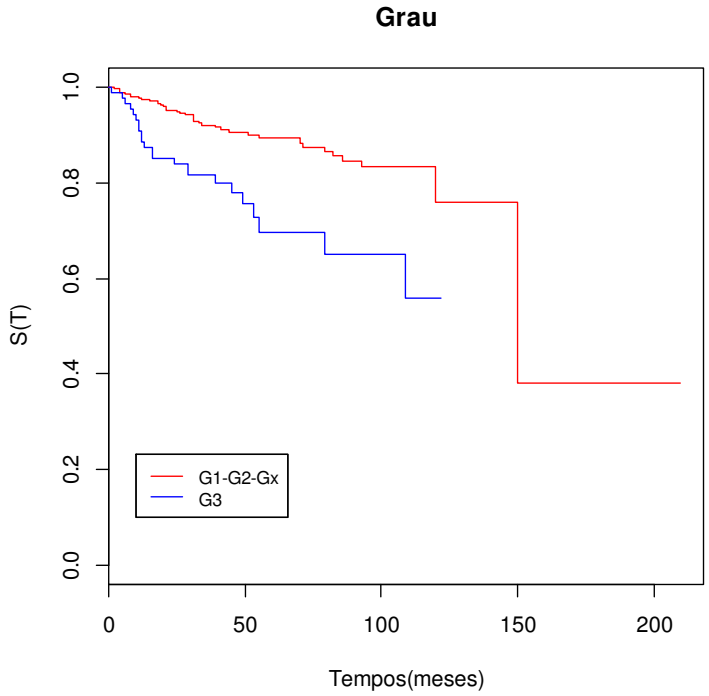



**Figura 34 - Curva de Kaplan- Meier para a variável recodificada grau**

Pela figura 35, podemos verificar que quanto maior for o tamanho do tumor maior será a probabilidade de a paciente ter recidiva.

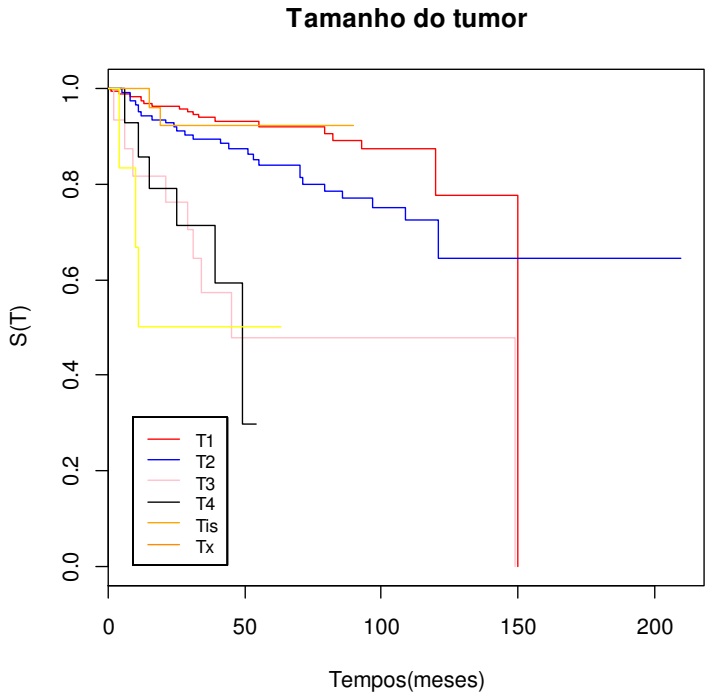



**Figura 35 - Curva de Kaplan-Meier para a variável tamanho do tumor**

Podemos também verificar que pacientes com tamanho de tumor nas categorias $T_1$, $T_2$, Tis, têm menor probabilidade de ter recidiva, comparativamente às pacientes que têm um tumor de tamanho na categoria $T_3$, $T_4$ ou $T_x$,

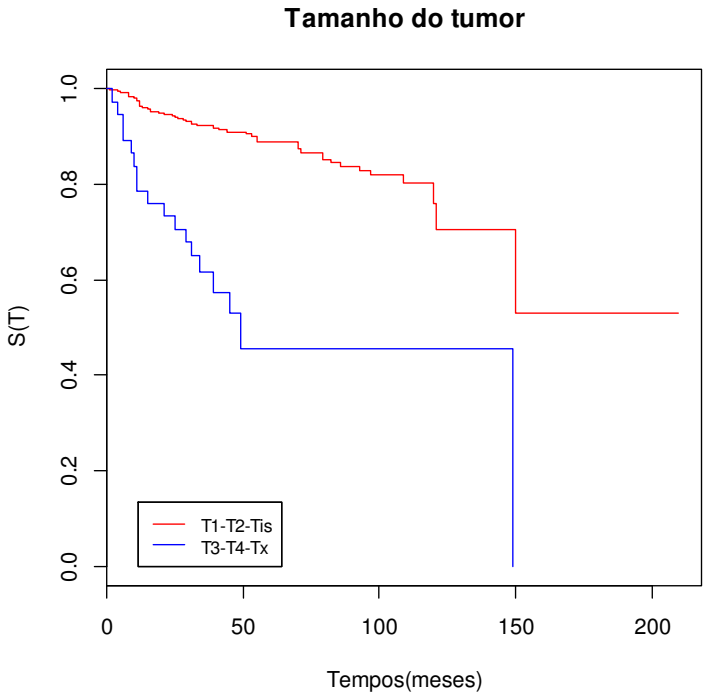



**Figura 36- Curva de Kaplan-Meier para a variável tamanho do tumor**

Após o teste $log - rank$ podemos verificar que existem diferenças significativas entre as curvas, com um valor de prova <0.001.

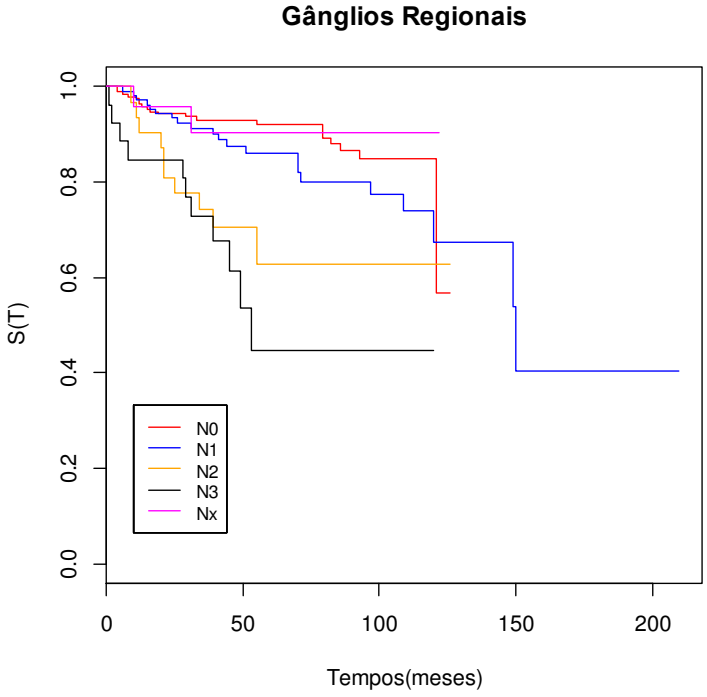



**Figura 37- Curva de Kaplan-Meier para a variável gânglios regionais**

Pelo gráfico anterior podemos verificar que as pacientes que possuem gânglios regionais nas categorias $N_x$, $N_0$, $N_1$ ou $N_2$ têm uma maior probabilidade de não terem recidiva.

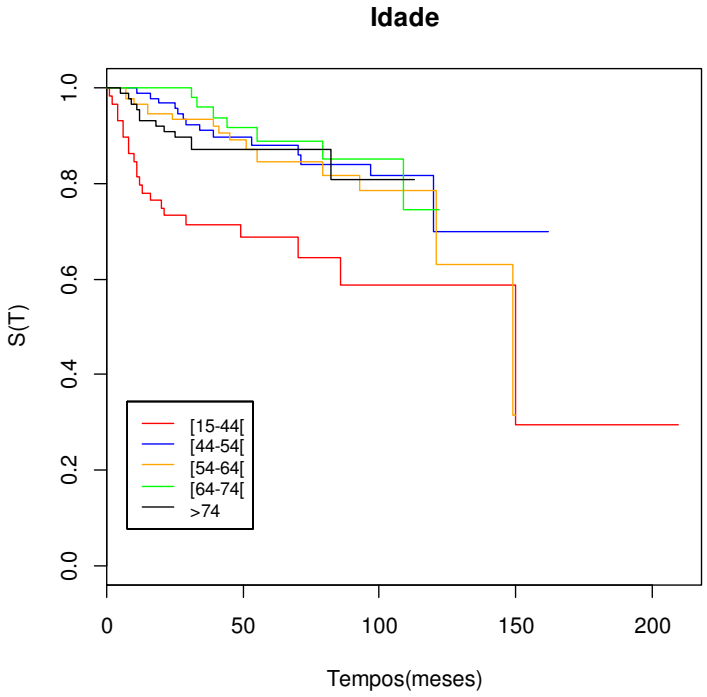



**Figura 38- Curva de Kaplan-Meier para a variável idade**

Como podemos verificar pela figura 38, existem diferenças entre as pacientes com idade ao diagnóstico igual ou mais de 44 anos comparativamente com as pacientes que tinham menos de 44 anos, ao diagnóstico. A figura mostra-nos que as pacientes com menos de 44 anos têm um menor risco de ter recidiva. O valor de prova < 0,001.

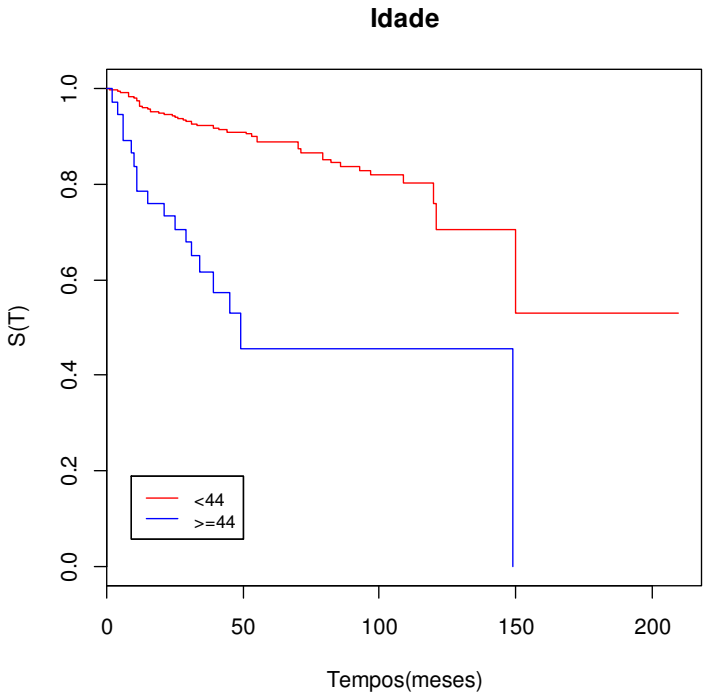



**Figura 39 - Gráfico de Kaplan-Meier para a variável recodificada idade**

Relativamente à hormonoterapia, podemos constatar pela figura 40, que as pacientes foram submetidas a este tratamento têm uma maior probabilidade de não ter recidiva comparativamente as que foram não foram

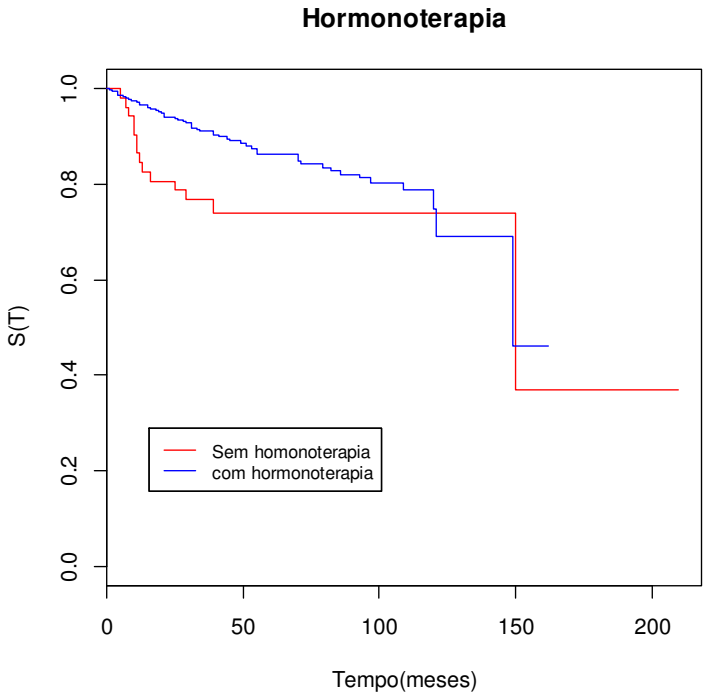



**Figura 40- Curva de Kaplan-Meier para a variável hormonoterapia**

## 5.2. Modelo de Cox

Ajustou-se inicialmente um modelo univariado a cada uma das variáveis com efeito significativo na probabilidade de recidiva.

Das variáveis em estudo, foram removidas as variáveis com mais de 50% de dados omissos.

Assim a, idade ao primeiro parto, relativa à paciente foi retirada. Das variáveis relativas ao tratamento que foram removidas foram: tratamento primário, tipo de cirurgia, hormonoterapia, quimioterapia, esvaziamento axilar e pesquisa de gânglios de sentinela.

## Modelo de Cox desde a data de início até à ocorrência da recidiva

### *Análise univariada*

**Tabela 37 - Resultados obtidos da análise univariada, pelo ajustamento do modelo de regressão de Cox**

| | $\beta$ | $e^{\beta}$ | Erro padrão | | Valor de prova |
|---|---|---|---|---|---|
| Estadio | 1.436 | 4.202 | 0.256 | 5.612 | <0.001 |
| Ki.67 | 0.365 | 1.441 | 0.269 | 1.357 | 0.175 |
| Gânglios | 1.375 | 3.956 | 0.336 | 4.098 | <0.001 |
| Imagens venosas | 1.240 | 3.454 | 0.380 | 3.259 | 0.001 |
| Imagens Linfáticas | 0.587 | 1.798 | 0.276 | 2.128 | 0.033 |
| RE | -0.675 | 0.509 | 0.317 | -2.125 | 0.034 |
| RP | -0.753 | 0.471 | 0.267 | -2.818 | 0.005 |
| TN | 1.855 | 6.389 | 0.370 | 5.009 | <0.001 |
| Grau | 0.985 | 2.664 | 0.273 | 3.604 | 0.000 |
| Idade | -1.039 | 0.354 | 0.266 | -3.9 | <0.001 |
| Tamanho do tumor | 1.756 | 5.788 | 0.288 | 6.107 | <0.001 |

Como podemos verificar pela tabela 37, todas as variáveis têm influência do risco de virem ter recidiva, exceto a variável Ki.67.

Após a análise univariada construímos um modelo de regressão de Cox múltipla, cujos resultados dos parâmetros estimados, respetivos erros padrões e valores de prova são apresentados na tabela 38.

### *Análise múltipla*

Após ser feita a seleção de variáveis pelo método $stepwise\ forward$ com a utilização do teste de razão de verosimilhanças é obtido um modelo de Cox a que correspondem os resultados da tabela 38.

**Tabela 38 - Resultados da análise múltipla do modelo de regressão de Cox**

| | $\beta$ | $e^{\beta}$ | Erro padrão | | Valor da prova |
|---|---|---|---|---|---|
| Triplo negativo | 2.091 | 8.091 | 0.393 | 5.327 | <0.001 |
| Tamanho | 1.390 | 4.015 | 0.309 | 4.501 | <0.001 |
| Gânglios | 1.202 | 3.327 | 0.374 | 3.216 | 0.001 |
| Idade | -0.926 | 0.396 | 0.286 | -3.238 | 0.001 |

Isoladamente todas as variáveis apresentaram-se significativas, no entanto, em conjunto só as variáveis: triplo negativo, tamanho do tumor, gânglios e idade é que têm influência do risco de virem a ter recidiva, no tempo até à recidiva.

Podemos também observar que:

- Uma paciente que apresente um tumor triplo negativo tem um risco de recidiva de cancro da mama aproximadamente 8 vezes superior do que uma paciente que apresente um tumor que não seja triplo negativo.
- Uma paciente que tenha um tumor na categoria $T_3$, $T_4$ ou $T_x$, tem um risco de recidiva de cancro da mama, 4 vezes maior do que uma paciente que tenha um tumor na categoria $T_1$ $T_2$ ou $T_{is}$.
- O risco de recidiva de cancro da mama de uma paciente em que os seus gânglios regionais se encontram na categoria $N_3$ é 3 vezes maior comparado à categoria de referência (categorias $N_x$, $N_0$, $N_1$ ou $N_2$)
- Relativamente à variável idade, o risco de ter recidiva do cancro da mama de uma paciente que tenha 44 ou mais anos é de aproximadamente 60% menor do que uma paciente que tenha menos de 44 anos.

### *Cox-Snell*

Através da análise dos resíduos de Cox-Snell é possível avaliar o ajustamento global do modelo. Na figura 41, observa-se a estimativa da função de risco dos resíduos de Cox-Snell, e ajustando uma reta com ordenada na origem nula e declive um, conclui-se que o modelo parece ser adequado.

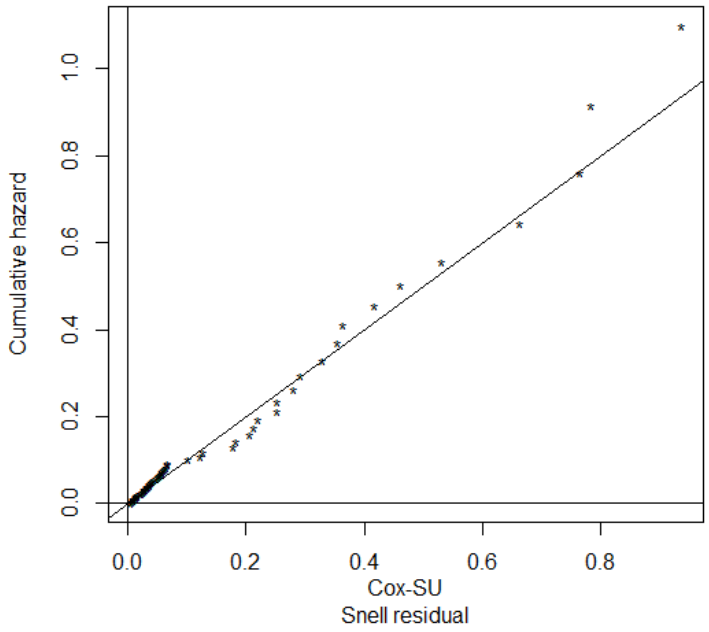



**Figura 41- Gráfico Cox- Snell**

Como referido anteriormente, a análise de proporcionalidade de riscos será apresentada através de um gráfico em que se representam os resíduos, os intervalos de confiança da curva de suavização *spline* dos resíduos e uma linha horizontal correspondente ao efeito constante da variável estimada pelo modelo.

Os resultados do teste de hipóteses de proporcionalidade são:

**Tabela 39- Testes de proporcionalidade dos riscos no modelo de Cox**

| | Chisq | Valor da prova |
|---|---|---|
| Triplo negativo | 0.068 | 0.795 |
| Tamanho do tumor | 0.096 | 0.757 |
| Gânglios regionais | 0.692 | 0.406 |
| Idade | 4.385 | 0.036 |
| Global | 4.836 | 0.306 |

É de salientar que o valor de prova para a variável idade é marginalmente significativo, no entanto foi incluída nos gráficos dos Resíduos Schoenfeld, para verificar a existência de proporcionalidade das funções de risco.

Constata-se que, na globalidade, o modelo não viola a hipótese de proporcionalidade das funções de risco, como podemos verificar pela tabela 39.

Em seguida apresentam-se os gráficos dos Resíduos Schoenfeld de cada variável do modelo

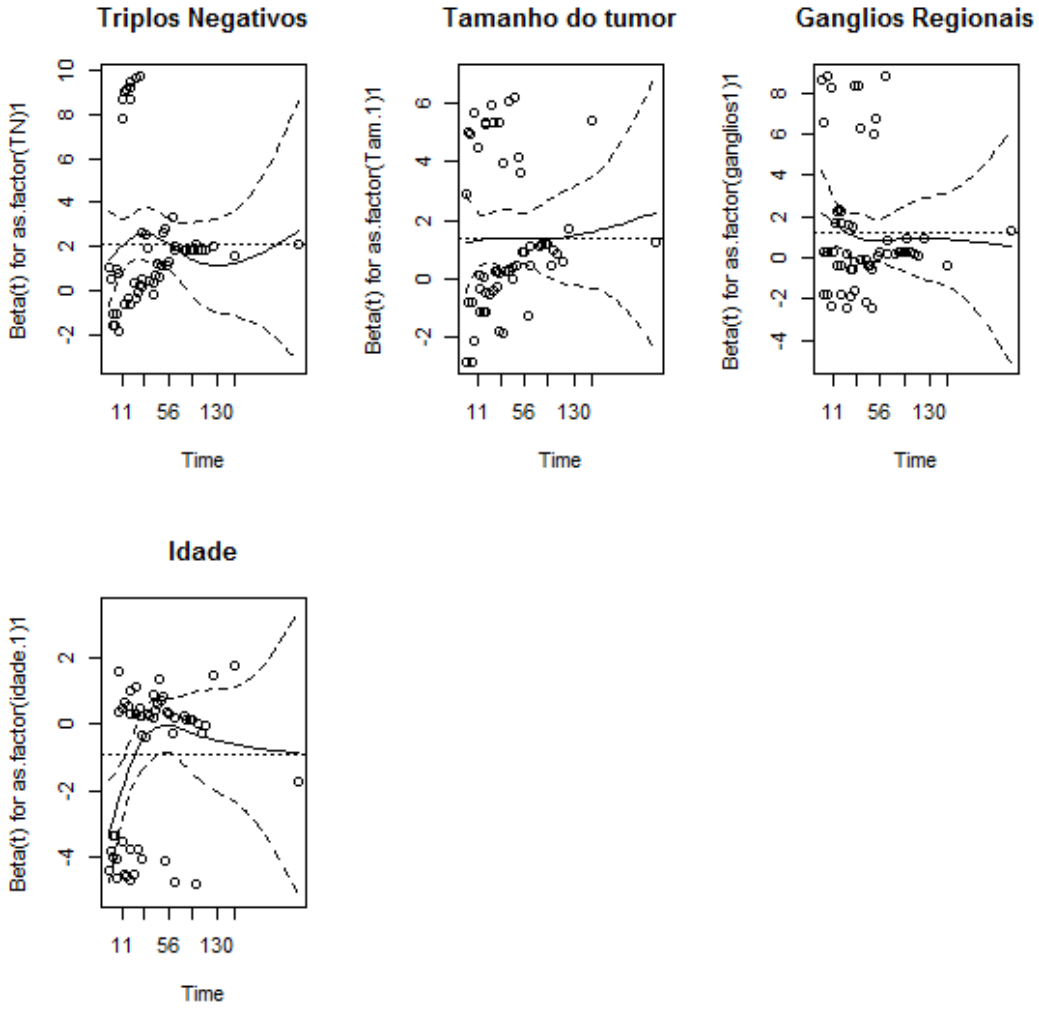



**Figura 42- Resíduos Schoenfeld**

A proporcionalidade dos riscos foi avaliada com base na representação gráfica dos resíduos de Schoenfeld (figura 42). Através destes gráficos é razoável presumir-se a existência de proporcionalidade das funções de risco, visto que não é encontrada uma tendência marcada nos resíduos Schoenfeld em função do tempo.

## Modelo de Cox data de tratamento até à recidiva

### *Análise univariada*

**Tabela 40- Resultados obtidos da análise univariada, pelo ajustamento do modelo de regressão de Cox**

| | $\beta$ | $e^{\beta}$ | Erro padrão | | Valor da prova |
|---|---|---|---|---|---|
| Estadio | 1,467 | 4,334 | 0,257 | 5,713 | <0.001 |
| Ki.67 | 0,371 | 1,450 | 0,277 | 1,366 | 0,172 |
| Gânglios | 1,390 | 4,014 | 0,337 | 4,128 | <0.001 |
| Imagens venosas | 1,248 | 3,483 | 0,382 | 3,271 | 0,001 |
| Imagens linfáticas | 0,621 | 1,861 | 0,278 | 2,236 | 0,025 |
| RE | -0,687 | 0,503 | 0,319 | -2,161 | 0,030 |
| RP | -0,77 | 0,463 | 0,268 | -2,878 | 0,004 |
| TN | 1,855 | 6,390 | 0,370 | 5,018 | <0.001 |
| Grau | 0,992 | 2,696 | 0,273 | 3,631 | <0.001 |
| Idade | -1,039 | 0,354 | 0,265 | -3,914 | <0.001 |
| Tamanho do tumor | 1,756 | 5,791 | 0,288 | 6,107 | <0.001 |

Isoladamente, todas as variáveis têm influência significativa exceto a variável Ki.67.

Depois de estimar o efeito que cada variável, por si só tem no tempo até a ocorrência de recidiva por cancro da mama, interessa agora construir um modelo de regressão de Cox multivariado.

### *Análise Múltipla*

Após ser feita a seleção de variáveis pelo método $stepwise\ forward$ com a utilização do teste de razão de verosimilhanças é obtido um modelo de Cox a que correspondem os resultados da tabela 41.

**Tabela 41 - Resultados da análise múltipla do modelo de regressão de Cox**

| | $\beta$ | $e^{\beta}$ | Erro padrão | | Valor da prova |
|---|---|---|---|---|---|
| Estadio | 1,027 | 2,792 | 0,351 | 2,927 | 0,003 |
| Triplo negativo | 1,941 | 6,969 | 0,400 | 4,857 | <0.001 |
| Tamanho do tumor | 0,859 | 2,362 | 0,400 | 2,151 | 0,032 |
| Idade | -0,976 | 0,377 | 0,280 | -3,484 | 0,001 |

Em conjunto com todas as variáveis, o estadio, o triplo negativo o tamanho do tumor e a idade passam a ter significado no tempo desde a data do tratamento até à recidiva.

Podemos constatar pela tabela que:

- Uma paciente com um tumor no estadio (III ou IV) tem um risco de recidiva de cancro da mama 2,8 vezes maior quando comparado à categoria de referencia (tumor no estadio (0, I ou II))
- Uma paciente que apresente um tumor triplo negativo tem um risco de recidiva de cancro da mama aproximadamente 7 vezes superior do que uma paciente que apresente um tumor que não seja triplo negativo.
- Uma paciente que tenha um tumor na categoria $T_3$, $T_4$ ou $T_x$, o seu risco de recidiva de cancro da mama2,4 vezes maior do que uma paciente que tenha um tumor na categoria $T_1$ $T_2$ ou $T_{is.}$
- Relativamente à variável idade, significa que o risco de ter recidiva do cancro da mamam de uma paciente que tem mais ou igual a 44 anos é de aproximadamente 60% menor do que uma paciente que tenha menos de 44 anos.

*Cox- Snell*

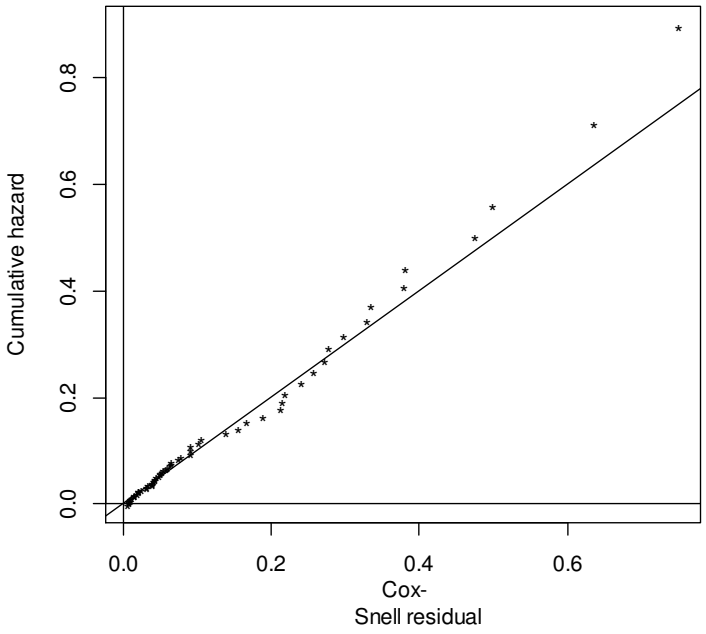



**Figura 43 - Gráfico de Cox- Snell**

Através da análise dos resíduos de Cox-Snell e possível avaliar o ajustamento global do modelo. Na figura 43, observa-se a estimativa da função de risco dos resíduos de Cox-Snell, e ajustando uma reta com ordenada na origem nula e declive um, conclui-se que o modelo parece ser adequado.

A verificação da hipótese de proporcionalidade de cada uma das variáveis deve ser feita considerando tanto o teste como o gráfico correspondente. Os resultados do teste de hipótese de proporcionalidade são:

**Tabela 42 - Testes de proporcionalidade dos riscos do modelo de Cox**

| | Chisq | Valor da prova |
|---|---|---|
| Estadio | 1,820 | 0,177 |
| TN | 0,223 | 0,637 |
| Tamanho do tumor | 1,204 | 0,273 |
| Idade | 3,044 | 0,081 |
| Global | 5,157 | 0,272 |

Constata-se na globalidade que o modelo não viola a hipótese de proporcionalidade das funções de risco, como podemos observar pelo valor do teste global (> 0,05) e dos testes de cada variável (> 0,30).

É de salientar que o valor de prova para a variável idade não é significativo, no entanto, foi incluída nos gráficos dos Resíduos Schoenfeld, para verificar a existência de proporcionalidade das funções de risco.

Em seguida apresentam-se os gráficos de Schoenfeld.

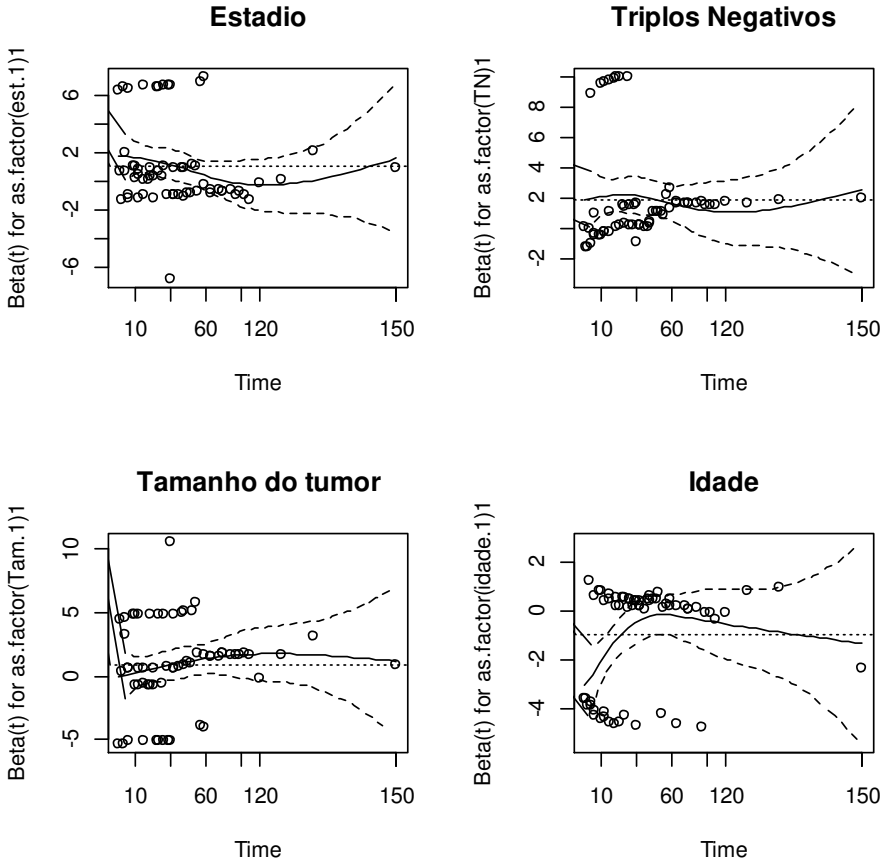



**Figura 44 - Resíduos padronizados de Schoenfeld no modelo de Cox**

A partir da figura 44, podemos observar em todos eles que o efeito parece ser diferente no fim da observação, no entanto, como existem poucas observações, não se valorizam estas variações. Além disso, os testes acima mencionados não rejeitam a hipótese de proporcionalidade dos riscos.

# Capítulo 6- Estudo Longitudinal de marcadores tumorais

## 6.1. Estudo longitudinal do marcador CA15.3

Para o marcador tumoral CA15.3 foram efetuadas 5166 medições desde o diagnóstico até ao final do estudo, em que o evento foi a recidiva. 36 mulheres tiveram cancro bilateral e 79 morreram.

### Tempo desde o diagnóstico até à data do teste – evento recidiva

A figura 45 representa a progressão no tempo, desde o tempo do diagnóstico até à data do teste, dos valores do marcador CA15.3. De notar que neste cenário, todas as medidas do marcador posterior à recidiva foram ignoradas.

As linhas a cinzento representam a progressão do marcador CA15.3 de cada paciente. Com esta análise permite-nos analisar a evolução do marcador tumoral desde o diagnóstico até a uma eventual recidiva.

A linha a tracejado corresponde ao valor de referência (log (37) (u/ml)) do marcador CA15.3. A linha mais saliente corresponde à tendência média da progressão do marcador tumoral, usando um modelo não paramétrico de suavização smooth.spline

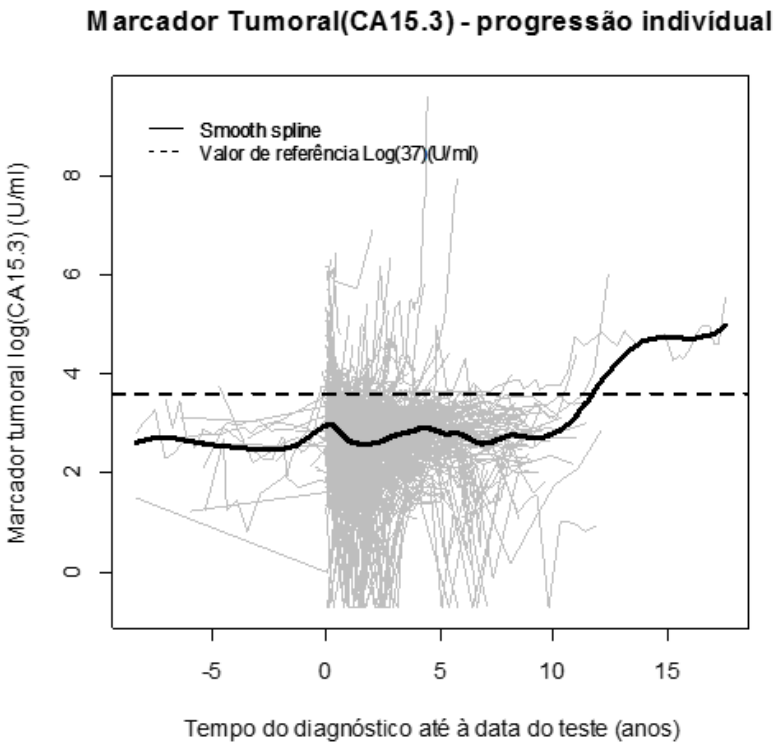



**Figura 45 - Progressão individual para o valor do marcador tumoral CA 15-3**

Foi realizada uma análise longitudinal a partir de todas as variáveis significativas nos modelos de sobrevivência, para as diferentes estruturas de correlação.

Iniciamos o nosso estudo com uma análise do modelo de regressão, que assume independência entre todas as observações, observações entre indivíduos e observações de um mesmo indivíduo, com o modelo saturado.

**Tabela 43- Valores obtidos pelo método dos mínimos quadrados – modelo saturado**

| | Estimativa | Erro padrão | Erro padrão robusto |
|---|---|---|---|
| Ordenada na origem | 2.501 | 0.120 | 0.300 |
| Tempo (anos) | -0.011 | 0.013 | 0.034 |
| Idade ao diagnóstico | 0.009 | 0.001 | 0.004 |
| Com cancro bilateral | -0.035 | 0.121 | 0.268 |
| Estadio – categoria III_IV | 0.474 | 0.060 | 0.240 |
| Grau na categoria G3 | -0.123 | 0.058 | 0.153 |
| Com presença de carcinoma associado | 0.003 | 0.041 | 0.114 |
| Com imagens de invasão vascular linfática | -0.007 | 0.047 | 0.117 |
| Com imagens de invasão vascular venosa | 0.360 | 0.094 | 0.277 |
| RE - positivo | -0.056 | 0.075 | 0.209 |
| RP - positivo | -0.016 | 0.061 | 0.165 |
| HER.2neu - positivo | -0.252 | 0.044 | 0.130 |
| Com triplo negativo | -0.544 | 0.141 | 0.327 |
| Ki.67- baixo | -0.251 | 0.043 | 0.111 |
| Com presença de gânglios regionais | -0.411 | 0.103 | 0.276 |

Os resultados obtidos pelo modelo de regressão independente, são os que estão representados na tabela43. Utilizando o método *stepwise*, descrito no capítulo 2, a partir do modelo saturado, fomos selecionar quais as variáveis que melhor descrevem a progressão do marcador CA15.3

Foi também utilizado o modelo longitudinal que explica a estrutura de correlação temporal exponencial, assim como o modelo longitudinal que explica a estrutura de correlação temporal gaussiana referenciado no capítulo 2, para verificar qual a estrutura com menor valor de likelihood.

**Tabela 44 - Valores dos diferentes modelos saturados**

| | OLS | Exponencial | Gaussiano |
|---|---|---|---|
| AIC | 3874.64 | 3259.89 | 3274.76 |
| LogLik | -1920.32 | -1610.94 | -1618.38 |

Como podemos verificar pela tabela 44 o modelo exponencial teve o menor valor likelihood, por isso escolhemos este modelo para descrever a progressão do marcador tumoral CA15-3.

**Tabela 45- Valores estimados para os modelos OLS e longitudinais**

| | OLS | | Exponencial | | Gaussiano | |
|---|---|---|---|---|---|---|
| | Est | Valor de Prova | Est | Valor de prova | Est | Valor de prova |
| Ordenada na origem | 2.300 | <0.001 | 2.275 | <0.001 | 2.264 | <0.001 |
| Tempo (anos) | 0.000 | 0.987 | 0.000 | 0.99 | 0.001 | 0.919 |
| Idade ao diagnóstico | 0.008 | 0.009 | 0.008 | 0.006 | 0.008 | 0.006 |
| Com imagens de invasão vascular venosa | 0.710 | <0.001 | 0.677 | <0.001 | 0.680 | <0.001 |
| Ki.67- baixo | -0.150 | 0.079 | -0.138 | 0.097 | -0.132 | 0.114 |

Partindo do modelo longitudinal com estrutura de correlação exponencial, retiraram-se uma a uma as covariáveis menos significativas, terminando com o modelo apresentado na tabela 45.Esta tabela apresenta, também, os parâmetros estimados (Est) dos modelos longitudinais comparados com as estimativas obtidas pelo modelo OLS e os respetivos valores de prova. Pela tabela percebe-se que as estimativas são semelhantes para os dois modelos mas existem ligeiras variações quanto à importância destes (valor de prova).

A parte fixa do modelo longitudinal que descreve a progressão média do marcador é composta pelas seguintes covariáveis: Tempo, idade ao diagnóstico, imagens de invasão vascular venosa (sim vs. não) e Ki.67 (alto vs. baixo).

O valor do log (CA15.3), à data do diagnóstico na média de idades com pacientes com imagem de invasão vascular venosa e Ki.67 baixo é de exp (2.275-0.01*0+0.0077*56+0.6769-0.1379) =25.66

Também podemos verificar que uma paciente com imagens de invasão vascular venosa tem um aumento no valor do marcador tumoral de exp (0.6769) =

1.9677quandocomparado com uma paciente que não tem imagens de invasão vascular venosa.

Uma paciente que tenha o antígeno Ki.67 baixo, o valor do marcador tumoral diminui exp (-0.1379) = 0.878 relativamente a uma paciente com o antígeno Ki.67 alto.

A figura 46 apresenta o variograma empírico e teórico para os diferentes modelos. Como podemos constatar o variograma do modelo exponencial é o que melhor se ajusta à curva empírica

A estrutura de correlação exponencial para descrever a variabilidade é dada por: $\rho(u) = \exp(-\frac{1}{0,55}|u|)$, onde $\boldsymbol{\delta}^2$ = 0.406, $\tilde{\imath}^2$ = 0,02 $\boldsymbol{\sigma^2} = 0,28$

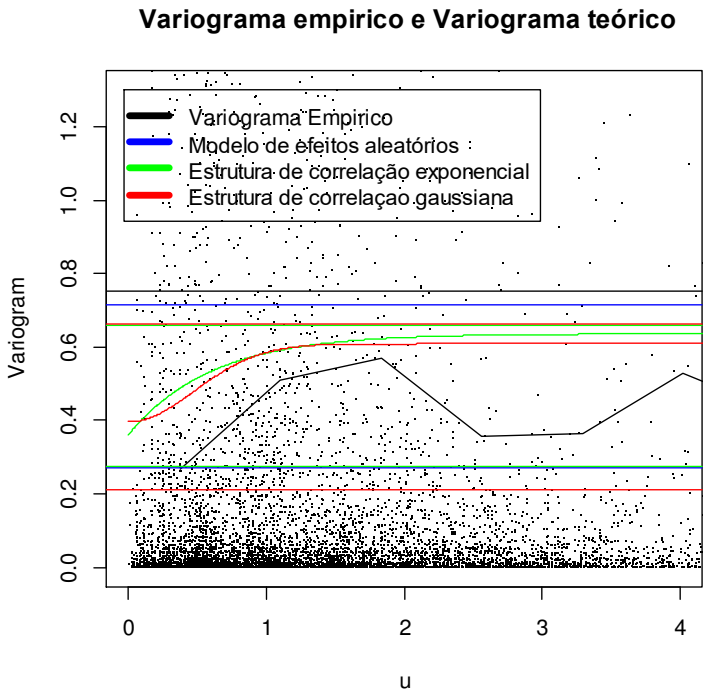



**Figura 46 - Sobreposição do variograma empírico e teórico do marcador CA15-3**

Ajustado o modelo, fomos verificar que efeito de cada uma das covariáveis na média. Para isso, sobrepusemos as respetivas médias teóricas estimadas pelo modelo sobre a progressão individual com a média empírica.

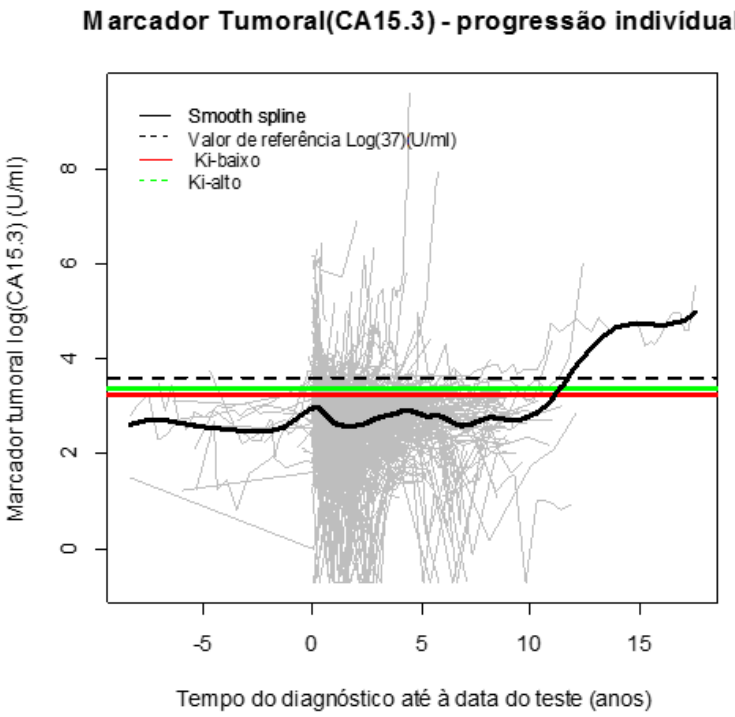



**Figura 47 -Sobreposição da progressão individual com a comparação das médias do Ki alto e baixo**

Pela figura 47 podemos observar que uma paciente com 56 anos de idade diagnosticada com invasão vascular venosa e um Ki baixo apresenta uma média do valor do marcador mais alto quando comparada com um Ki alto

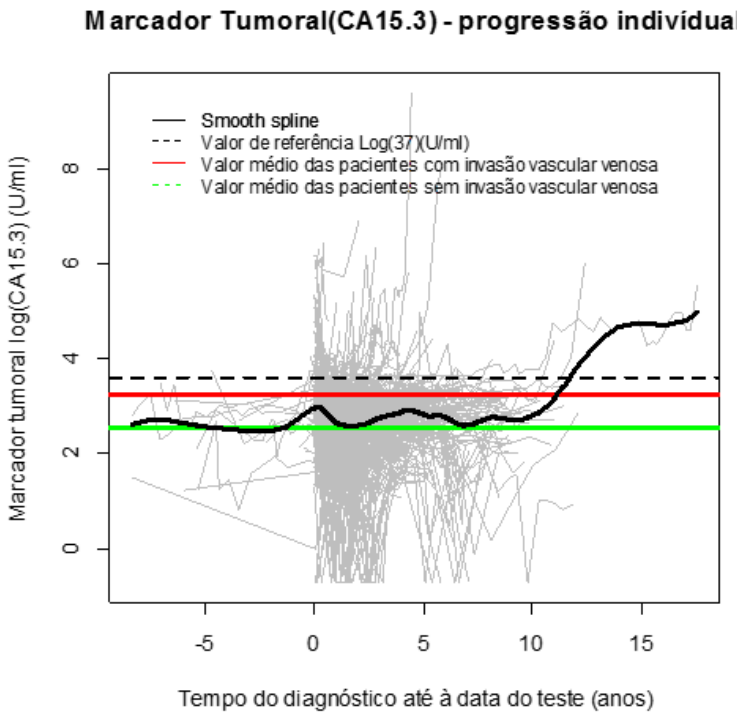



**Figura 48 - Sobreposição da progressão individual com a comparação das médias das pacientes com e sem invasão vascular venosa**

Pela figura 48, podemos observar que uma paciente com 56 anos de idade diagnosticada com invasão vascular venosa e um Ki alto apresenta uma média do valor do marcador mais alto quando comparada com uma paciente com Ki alto mas sem invasão vascular venosa.

## Tempo desde a recidiva até à data do teste

A figura 49 representa a progressão no tempo, desde o tempo da recidiva até à data do teste, dos valores do marcador CA15.3. De notar aqui, que esta análise é feita apenas para os indivíduos que tiveram recidiva. Por isso a interpretação é, com este modelo conseguimos entender a evolução do marcador, para o subconjunto de indivíduos com recidiva, antes e depois da recidiva.

As linhas a cinzento representam a progressão do marcador CA15.3 de cada paciente.

A linha a tracejado corresponde ao valor de referência (log (37) (u/ml)) do marcador CA15.3. A linha mais saliente corresponde à tendência média da progressão do marcador tumoral.

Neste caso em particular podemos verificar, pela progressão individual que o valor médio (*smoth spline*) do marcador tumoral aumenta a partir da data da recidiva, ou seja tempo = 0.

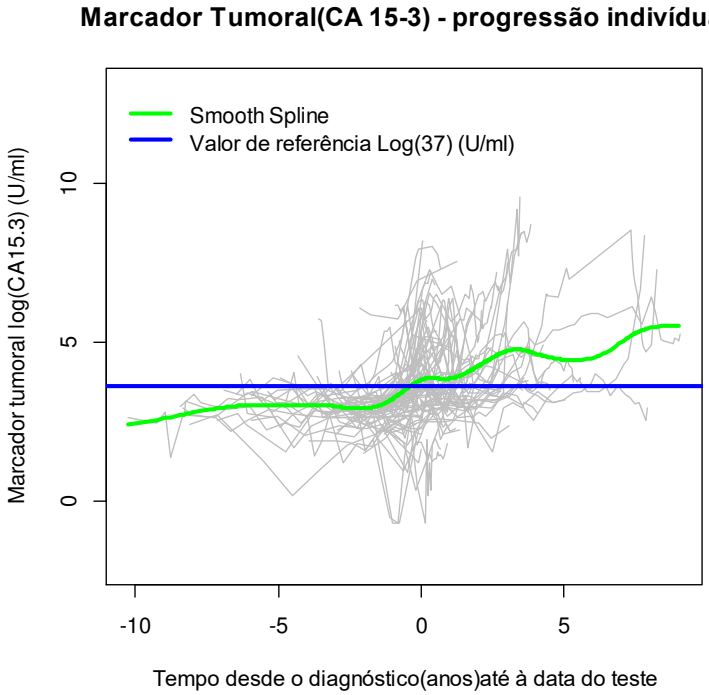



**Figura 49 -Progressão individual para o valor do marcador tumoral CA 15-3**

Foi realizada uma análise longitudinal a partir de todas as variáveis significativas nos modelos de sobrevivência, para as diferentes estruturas de correlação.

Iniciamos o nosso estudo com uma análise do modelo de regressão, que assume independência entre todas as observações, observações entre indivíduos e observações de um mesmo indivíduo, com o modelo saturado.

**Tabela 46- Valores obtidos pelo método dos mínimos quadrados – modelo saturado**

| | Estimativa | Erro padrão | Erro padrão robusto |
|---|---|---|---|
| Ordenada na origem | 1.289 | 0.585 | 1.273 |
| Tempo (anos) | 0.433 | 0.047 | 0.111 |
| Idade ao diagnóstico | 0.043 | 0.007 | 0.015 |
| Com cancro bilateral | -0.511 | 0.493 | 0.856 |
| Estadio na categoria III_IV | -0.645 | 0.271 | 0.502 |
| Grau na categoria G3 | 0.169 | 0.238 | 0.553 |
| Com presença de carcinoma associado | 0.631 | 0.245 | 0.359 |
| Com Imagens de invasão vascular linfática | 0.497 | 0.161 | 0.400 |
| Com Imagens de invasão vascular venosa | 2.175 | 0.275 | 0.658 |
| RE - positivo | 0.050 | 0.343 | 0.846 |
| RP - positivo | -0.444 | 0.246 | 0.558 |
| HER.2neu - positivo | -0.164 | 0.245 | 0.523 |
| Com triplos negativos | -1.403 | 0.416 | 0.970 |
| Ki.67 - baixo | 0.027 | 0.196 | 0.407 |
| Com presença de gânglios regionais | -0.246 | 0.338 | 0.676 |

Os resultados obtidos pelo modelo de regressão independente, são os que estão representados na tabela 46

Utilizando o método *stepwise*, descrito no capítulo 2, a partir do modelo saturado, fomos selecionar quais as variáveis que melhor descrevem a progressão do marcador CA15.3

Foi também utilizado o modelo longitudinal que explica a estrutura de correlação exponencial, assim como o modelo longitudinal que explica a estrutura de

correlação gaussiana referenciado no capítulo 2, para verificar qual a estrutura com menor valor de likelihood.

**Tabela 47 - Valores para os diferentes modelos**

| | OLS | Exponencial | Gaussiano |
|---|---|---|---|
| AIC | 1086.27 | 743.737 | 760.366 |
| LogLik | -525.13 | -351.86 | -360.183 |

Como podemos verificar pela tabela 47, o modelo exponencial apresenta o menor valor de Log-Likelihood, por isso escolhemos este modelo para descrever a progressão do marcador tumoral CA15-3.

A tabela 48 apresenta os valores dos parâmetros estimados (Est) para os modelos longitudinais comparados com o modelo OLS.

**Tabela 48 - Valores estimados para os modelos OLS e longitudinal**

| | OLS | | Exponencial | | Gaussiano | |
|---|---|---|---|---|---|---|
| | Est | Valor de prova | Est | Valor de prova | Est | Valor de prova |
| Ordenada na origem | 3.454 | <0.001 | 3.57 | <0.001 | 3.336 | <0.001 |
| Time .bf | 0.124 | <0.001 | 0.145 | <0.001 | 0.116 | <0.001 |
| Time.af | 0.292 | <0.001 | 0.263 | <0.001 | 0.285 | <0.001 |
| Com imagem de invasão vascular venosa | 1.287 | <0.001 | 1.363 | <0.001 | 1.396 | <0.001 |

Partindo do modelo longitudinal com estrutura de correlação exponencial, retiraram-se uma a uma as covariáveis menos significativas, terminando com o modelo apresentado na tabela 48. Esta tabela apresenta, também, os parâmetros estimados (Est) dos modelos longitudinais comparados com as estimativas obtidas pelo modelo OLS e os respetivos valores de prova. Pela tabela percebe-se que as estimativas são semelhantes para os dois modelos.

A partir do modelo exponencial ajustou-se um modelo para a média, utilizando um ponto de mudança no tempo = 0.

O ponto de mudança é o momento em que há uma alteração no declive da progressão média da variável resposta.

$$\mu_{ij} = \begin{cases} X_{ij}\,\beta + \propto_1 t_{ij}\; se\; t_{ij} < 0 \\ X_{ij}\,\beta + \propto_2 \left(t_{ij} - \delta\right) se\; t_{ij} \geq 0 \end{cases}$$

$\propto_1$e $\propto_2$ são os coeficientes que representam o declive antes do tempo zero, ou seja, antes da recidiva, e depois do tempo zero, depois da recidiva, respetivamente.

A parte fixa do modelo longitudinal que descreve a progressão média do marcador é composta pelas seguintes covariáveis: Tempo (time.bf - tempo após a recidiva- time.af - tempo antes da recidiva) e imagem de invasão venosa. O valor do log de CA15-3 à data da recidiva é de exp (3.57+0.145*0+0.263*0+1.363) = 138.79

Podemos constar pela tabela que, uma paciente que tenha invasão vascular venosa tem um aumento no valor do log do marcador tumoral de exp (1.363) = 3.9 quando comparado com uma paciente que não tenha invasão vascular venosa.

A figura 50 apresenta o variograma empírico e teórico para os diferentes modelos. Como podemos constatar o variograma do modelo exponencial é o que melhor se ajusta à curva empírica.

A estrutura de correlação exponencial para descrever a variabilidade é dada por: $\rho(u) = \exp(-\frac{1}{6{,}575}|u|)$, onde $\boldsymbol{\delta}^{\,2}$ = 0,135, ĩ² = 8,95e-09 $\boldsymbol{\sigma^2}$ $= 1{,}48$

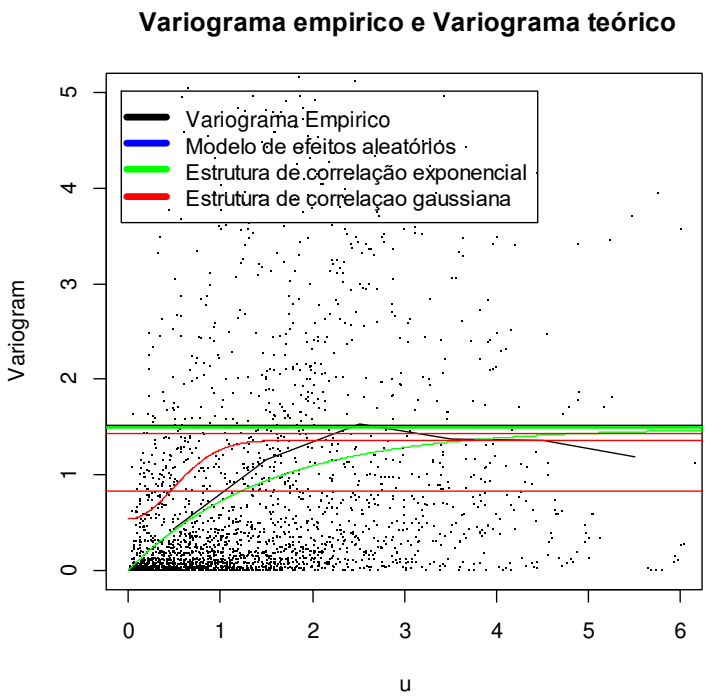



**Figura 50 - Sobreposição do variograma empírico e teórico do marcador CA15-3**

Ajustado o modelo, fomos verificar que efeito de cada uma das covariáveis na média. Para isso, sobrepusemos as respetivas médias teóricas estimadas pelo modelo sobre a progressão individual com a média empírica.

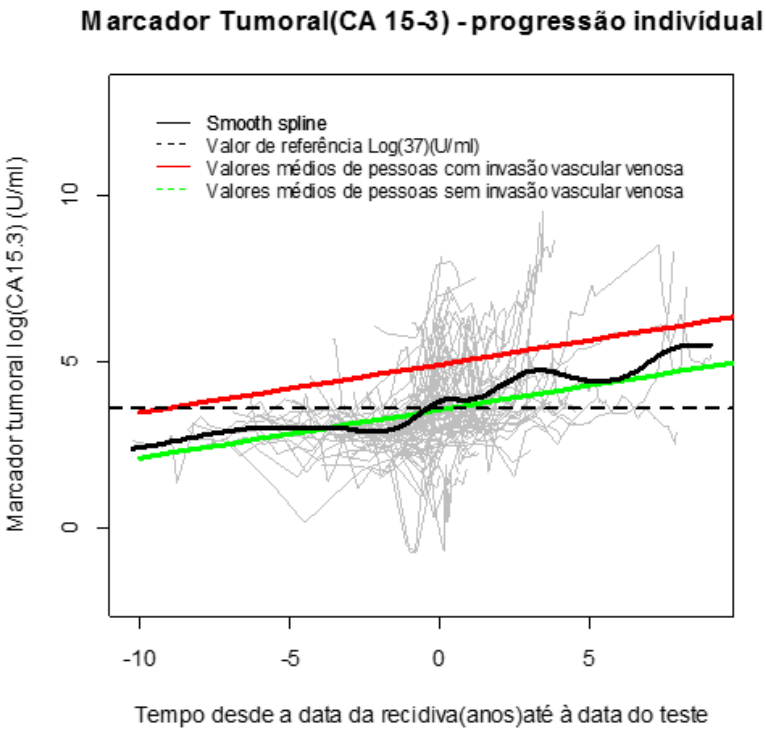



**Figura 51 - Sobreposição da progressão individual com a comparação das médias das pacientes com e sem invasão vascular venosa**

Pela figura 51 podemos observar que uma paciente, com imagem de invasão venosa apresenta uma média do valor do marcador, mais alto quando comparada com uma paciente sem imagem de invasão venosa.

## 6.2. Análise longitudinal do marcador CEA

### Tempo desde o diagnóstico até à data do teste- evento recidiva

A figura 52 apresenta a progressão no tempo desde o tempo do diagnóstico até à data do teste, dos valores do marcador CEA. De notar que neste cenário, todas as medidas do marcador posterior à recidiva foram ignoradas.

As linhas a cinzento representam a progressão do marcador CEA de cada paciente. Com esta análise permite-nos analisar a evolução do marcador tumoral desde o diagnóstico até a uma eventual recidiva.

A linha a tracejado corresponde ao valor de referência (log (5) (u/ml)) do marcador CEA. A linha mais saliente corresponde à tendência média da progressão do marcador tumoral, usando um modelo não paramétrico de suavização smooth.spline

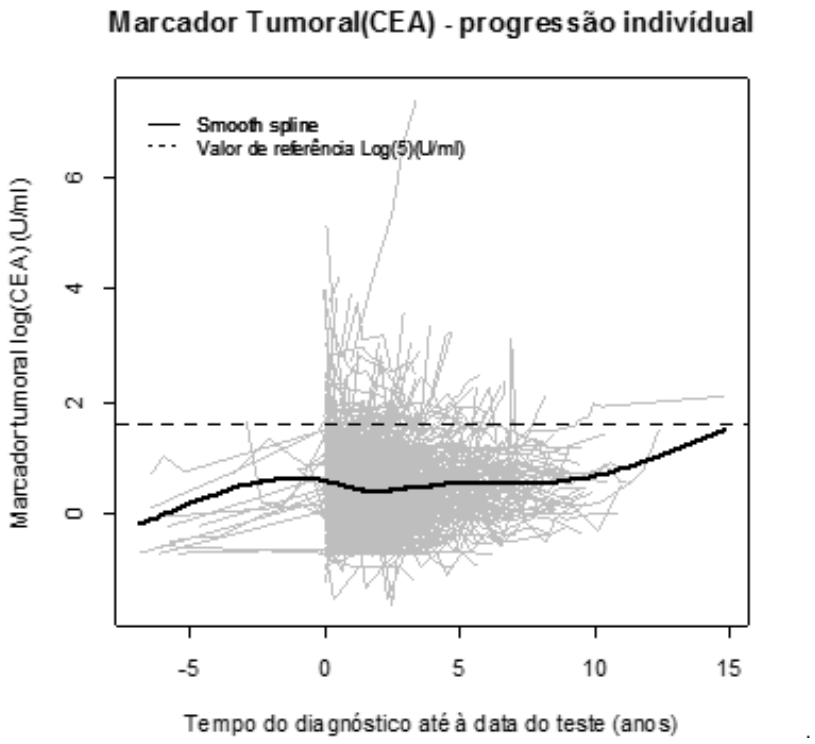



**Figura 52 -Progressão individual para o valor do marcador tumoral CEA**

Foi realizada uma análise longitudinal a partir de todas as variáveis significativas nos modelos de sobrevivência, para as diferentes estruturas de correlação.

Iniciamos o nosso estudo com uma análise do modelo de regressão, que assume independência entre todas as observações, observações entre indivíduos e observações de um mesmo indivíduo, com o modelo saturado.

Os resultados obtidos foram os obtidos na tabela 49.

**Tabela 49 - Valores obtidos pelo método dos mínimos quadrados – modelo saturado**

| | Estimativa | Erro padrão | Erro padrão robusto |
|---|---|---|---|
| Ordenada na origem | -0.642 | 0.122 | 0.278 |
| Tempo (anos) | -0.033 | 0.015 | 0.023 |
| Idade ao diagnóstico | 0.013 | 0.002 | 0.004 |
| Com cancro bilateral | 0.005 | 0.196 | 0.106 |
| Estadio na categoria III- IV | 0.016 | 0.075 | 0.174 |
| Grau na categoria $G_3$ | 0.138 | 0.065 | 0.123 |
| Com presença de carcinoma associado | -0.146 | 0.047 | 0.112 |
| Com imagens de invasão vascular linfática | 0.067 | 0.051 | 0.104 |
| Com imagens de invasão vascular venosa | -0.162 | 0.092 | 0.207 |
| RE - positivo | 0.459 | 0.082 | 0.214 |
| RP - positivo | 0.080 | 0.072 | 0.242 |
| HER.2neu - positivo | 0.009 | 0.048 | 0.107 |
| Com triplo negativo | 0.674 | 0.168 | 0.666 |
| Ki.67- baixo | 0.006 | 0.048 | 0.111 |
| Com presença de gânglios regionais | -0.091 | 0.119 | 0.287 |

Utilizando o método *stepwise*, descrito no capítulo 2, a partir do modelo saturado, fomos selecionar quais as variáveis que melhor descrevem a progressão do marcador CEA.

Foi também utilizado o modelo longitudinal que explica a estrutura de correlação exponencial, assim como o modelo longitudinal que explica a estrutura de correlação gaussiana referenciado no capítulo 2, para verificar qual a estrutura com menor valor de likelihood.

**Tabela 50 - Valores obtidos pelos diferentes modelos**

| | OLS | Exponencial | Gaussiano |
|---|---|---|---|
| AIC | 1047.49 | 907.81 | 923.44 |
| LogLik | -506.74 | -434.91 | -442.7 |

Como podemos constatar pelas tabelas anteriores o modelo que tem menor valor de Likelihood é o modelo exponencial, por isso escolhemos este modelo para descrever a progressão do marcador tumoral CEA.

| | OLS | | Exponencial | | Gaussiano | |
|---|---|---|---|---|---|---|
| | Est | Valor de prova | Est | Valor de prova | Est | Valor de prova |
| Ordenada na origem | -0.545 | <0.001 | -0.5198 | <0.001 | -0.529 | <0.001 |
| Anos | 0.054 | <0.001 | 0.0385 | <0.001 | 0.044 | <0.001 |
| Idade ao diagnóstico | 0.015 | <0.001 | 0.015 | <0.001 | 0.015 | <0.001 |
| Estadio na categoria III-IV | 0.2978 | <0.001 | 0.3019 | <0.001 | 0.307 | <0.001 |

**Tabela 51 - Valores estimados para os modelos OLS e longitudinal**

Partindo do modelo longitudinal com estrutura de correlação exponencial, retiraram-se uma a uma as covariáveis menos significativas, terminando com o modelo apresentado na tabela 54. Esta tabela apresenta, também, os parâmetros estimados (Est) dos modelos longitudinais comparados com as estimativas obtidas pelo modelo OLS e os respetivos valores de prova. Pela tabela percebe-se que as estimativas são semelhantes para os dois modelos

A parte fixa do modelo longitudinal que descreve a progressão média do marcador CEA, é composta pelas seguintes covariáveis: anos, idade ao diagnóstico e estadio do tumor.

O valor do log de CEA à data do diagnóstico com pacientes no estadio (III ou IV) é de exp ( -0.5198+ 0.0385* 0+0.015*56+0.3019) = 1.8629

Podemos verificar pela tabela que uma paciente com o estadio na categoria (III ou IV) tem um aumento no valor do marcador tumoral de exp (0.3019) = 1,352 quando comparado com uma paciente com o estadio na categoria (0- I ou II).

A figura 53 apresenta o variograma empírico e teórico para os diferentes modelos. Como podemos constatar o variograma exponencial é o que melhor se ajusta à curva empírica.

A estrutura que melhor representa a variabilidade é a que incorpora efeitos aleatórios.

A estrutura de correlação exponencial para descrever a variabilidade é dada por: $\rho(u) = \exp(-\frac{1}{0.274}|u|)$, onde $\boldsymbol{\delta}^2 = 0.274$, $\tilde{\imath}^2 = 0.0392\boldsymbol{\sigma}^2 = 0.17514$

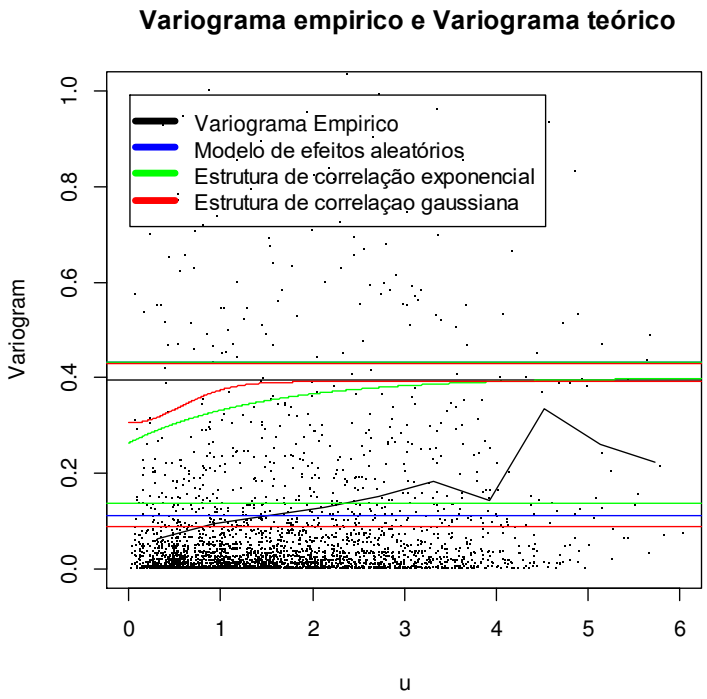



**Figura 53- Sobreposição do variograma empírico e teórico para o marcador CEA**

Ajustado o modelo, fomos verificar qual o efeito de cada uma das covariáveis na média. Para isso, sobrepusemos as respetivas médias teóricas estimadas pelo modelo sobre a progressão individual com a média empírica

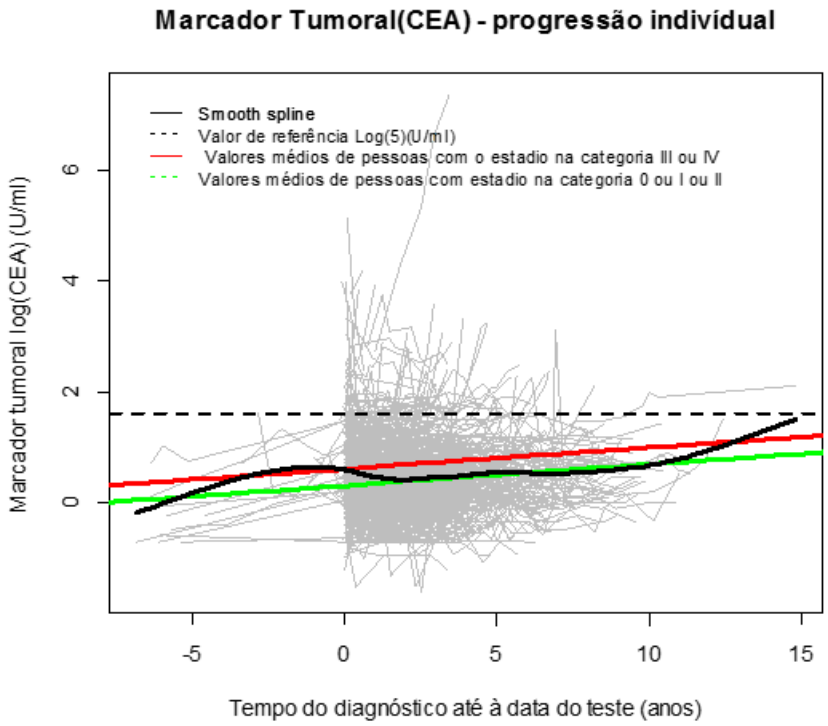



**Figura 54 - Sobreposição da progressão individual com a comparação das médias das pacientes nos diferentes estadios categorizados**

Pela figura 54 podemos observar que uma paciente diagnosticada no estadio III ou IV apresenta uma média do valor do marcador, mais alta quando comparada com uma paciente diagnosticada com estadio na categoria 0, I ou II.

## Tempo desde a recidiva até à data do teste

A figura 55 representa a progressão no tempo, desde o tempo da recidiva até à data do teste, dos valores do marcador CEA. De notar aqui, que esta análise é feita apenas para os indivíduos que tiveram recidiva. Por isso a interpretação é, com este modelo conseguimos entender a evolução do marcador, para o subconjunto de indivíduos com recidiva, antes e depois da recidiva.

As linhas a cinzento representam a progressão do marcador de cada paciente. A linha a azul corresponde ao valor de referência do marcador CEA (log (5) u/ml). A linha a verde corresponde à tendência média da progressão do marcador tumoral. Neste caso em particular podemos verificar, pela progressão individual que o

valor médio (*smoth spline*) do marcador tumoral aumenta a partir da data da recidiva, ou seja tempo = 0.

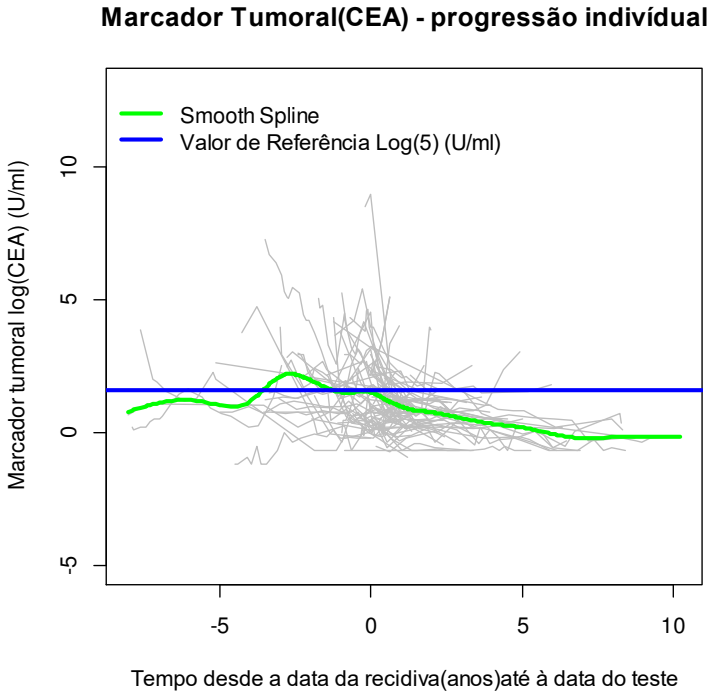



**Figura 55 -Progressão individual para o valor do marcador tumoral CEA**

Foi realizada uma análise longitudinal a partir de todas as variáveis significativas nos modelos de sobrevivência, para as diferentes estruturas de correlação.

Iniciamos o nosso estudo com uma análise do modelo de regressão, que assume independência entre todas as observações, observações entre indivíduos e observações de um mesmo indivíduo, com o modelo saturado.

Os resultados encontram-se na tabela 52.

**Tabela 52 - Valores obtidos pelo método dos mínimos quadrados – modelo saturado**

| | Estimativa | Erro padrão | Erro padrão robusto |
|---|---|---|---|
| Ordenada na origem | 2.605 | 1.362 | 1.854 |
| Tempo (anos) | -0.341 | 0.090 | 0.106 |
| Idade ao diagnóstico | -0.066 | 0.025 | 0.043 |
| Com cancro Bilateral | -3.571 | 1.508 | 1.912 |
| Estadio1 na categoria III_IV | 4.055 | 1.055 | 1.526 |
| Grau na categoria $G_3$ | 1.604 | 0.696 | 0.836 |
| Com presença de carcinoma associado | 0.186 | 0.436 | 0.357 |
| Com imagens de invasão vascular linfática | -2.607 | 0.564 | 1.048 |
| Com imagens de invasão vascular venosa | -0.377 | 0.940 | 0.753 |
| RE - positivo | -0.126 | 1.373 | 1.399 |
| RP - positivo | 0.653 | 0.755 | 0.983 |
| HER.2neu - positivo | 3.292 | 1.016 | 1.386 |
| Com Triplo negativo | 0.519 | 1.850 | 2.054 |
| Ki.67 - baixo | -0.282 | 0.409 | 0.558 |
| Com presença de gânglios regionais | -3.583 | 1.186 | 1.428 |

Utilizando o método *stepwise*, descrito no capítulo 2, a partir do modelo saturado, fomos selecionar quais as variáveis que melhor descrevem a progressão do marcador CEA.

Foi também utilizado o modelo longitudinal que explica a estrutura de correlação exponencial, assim como o modelo longitudinal que explica a estrutura de correlação gaussiana referenciado no capítulo 2, para verificar qual a estrutura com menor valor de likelihood.

**Tabela 53- Valores obtidos pelos diferentes modelos**

| | OLS | Exponencial | Gaussiano |
|---|---|---|---|
| AIC | 340.17 | 293.92 | 271.04 |
| LogLik | -152.08 | -126.96 | -115.52 |

Como podemos verificar pela tabela 53, o modelo Gaussiano teve o menor valor de AIC, por isso escolhemos este modelo para descrever a progressão do marcador tumoral.

| | OLS | | Exponencial | | Gaussiano | |
|---|---|---|---|---|---|---|
| | Est | Valor de prova | Est | Valor de prova | Est | Valor de prova |
| Ordenada na origem | 1.12 | <0.001 | 1.14 | <0.001 | 1.111 | <0.001 |
| Time.af | -0.22 | <0.001 | -0.23 | <0.001 | -0.23 | <0.001 |
| Time.bf | -0.33 | <0.001 | -0.18 | 0.005 | -0.23 | 0.001 |
| Bilateral | -0.94 | <0.001 | -0.74 | 0.03 | -0.783 | 0.002 |
| Estadio na categoria (III-IV) | 0.59 | <0.001 | 0.62 | 0.001 | 0.582 | <0.001 |
| Grau $G_3$ | 0.51 | <0.001 | 0.47 | 0.01 | 0.448 | <0.001 |
| Com triplo negativo | -1.25 | <0.001 | -1.21 | <0.001 | -1.19 | <0.001 |

**Tabela 54- Valores estimados para os modelos OLS e longitudinal**

Partindo do modelo longitudinal com estrutura de correlação exponencial, retiraram-se uma a uma as covariáveis menos significativas, terminando com o modelo apresentado na tabela 54. Esta tabela apresenta, também, os parâmetros estimados (Est) dos modelos longitudinais comparados com as estimativas obtidas pelo modelo OLS e os respetivos valores de prova. Pela tabela percebe-se que as estimativas são semelhantes para os dois modelos, mas existem ligeiras variações quanto à importância destes (valor de prova).

A partir do modelo exponencial ajustou-se um modelo para a média, utilizando um ponto de mudança no tempo = 0.

O ponto de mudança é o momento em que há uma alteração no declive da progressão média da variável resposta.

$$\mu_{ij} = \begin{cases} X_{ij}\,\beta + \propto_1 t_{ij}\; se\; t_{ij} < 0 \\ X_{ij}\,\beta + \propto_2 \left(t_{ij} - \delta\right) se\; t_{ij} \geq 0 \end{cases}$$

$\propto_1$e $\propto_2$ são os coeficientes que representam o declive antes do tempo zero, ou seja, antes da recidiva, e depois do tempo zero, depois da recidiva, respetivamente.

O valor do log de CEA à data do diagnóstico é de exp (1,11-0.23* 0- 0,23 * 0 - 0,783+0,582+0,448-1,19) = 1,18

Podemos verificar que uma paciente com um tumor bilateral tem uma diminuição no marcador tumoral de exp (-0,783) = 0.45 quando comparado com uma paciente que não possui tumor bilateral.

Uma paciente com um tumor no estadio III-IV tem um aumento no marcador tumoral de exp (0,59) = 1.80 quando comparado com uma paciente que tem um tumor no estadio 0-I-II.

Relativamente ao grau, podemos verificar que uma paciente com um tumor no grau $G_3$, tem um aumento no marcador tumoral de exp (0,448) = 1,565, quando comparado com uma paciente que tem um tumor no grau $G_1$, $G_2$, Gx.

Uma paciente com triplo negativo tem um aumento no marcador tumoral de exp (-1.19) = 0.30 quando comparado com uma paciente que não tem triplo negativo.

A estrutura que melhor representa a variabilidade é a que incorpora efeitos aleatórios.

A estrutura de correlação gaussiana para descrever a variabilidade é dada por:
$\rho(u) = \exp(-\frac{1}{0.672}u^2)$, onde $\boldsymbol{\delta}^2$ = 0.20, ĩ² = 0,08 $\boldsymbol{\sigma^2}$ = 0,77

A figura 56 apresenta o variograma empírico e teórico para os diferentes modelos. Como podemos verificar o variograma do modelo com estrutura de correlação temporal gaussiano é o melhor que se ajusta à curva empírica.

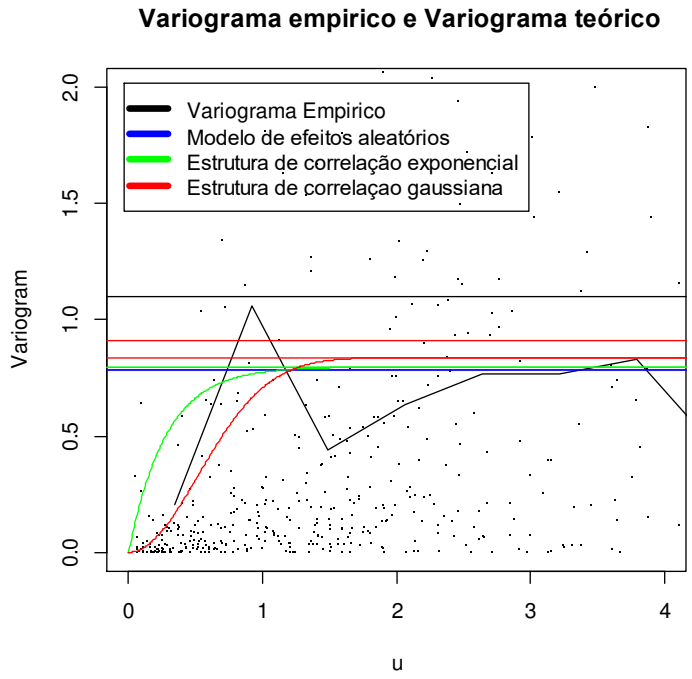



**Figura 56 - Sobreposição do variograma empírico e teórico do marcador CEA**

Ajustado o modelo, fomos verificar que efeito de cada uma das covariáveis na média. Para isso, sobrepusemos as respetivas médias teóricas estimadas pelo modelo sobre a progressão individual com a média empírica.

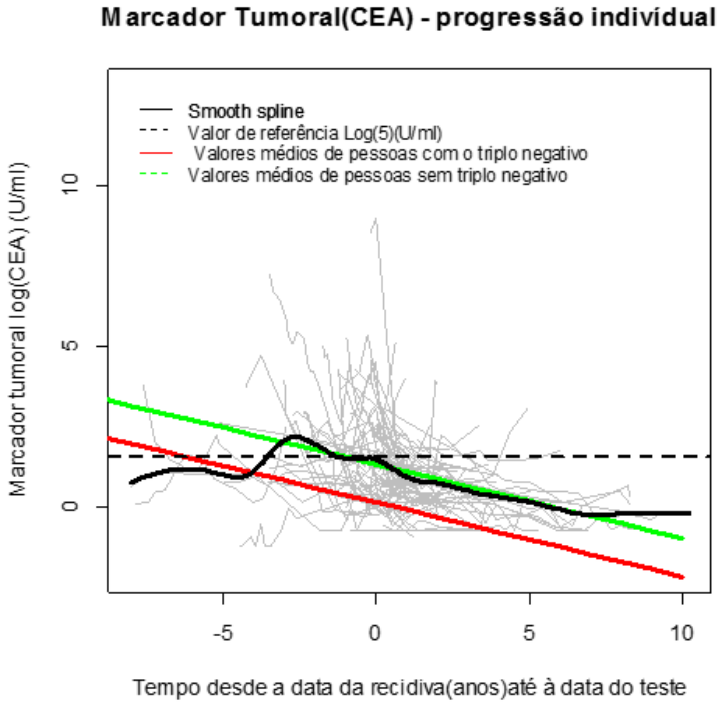



**Figura 57 - Sobreposição da progressão individual com a comparação das médias das pacientes com triplo negativo**

Pela figura 57 podemos observar que uma paciente com triplo negativo apresenta uma média do valor do marcador, mais alto quando comparada com uma paciente sem a presença de triplo negativo.

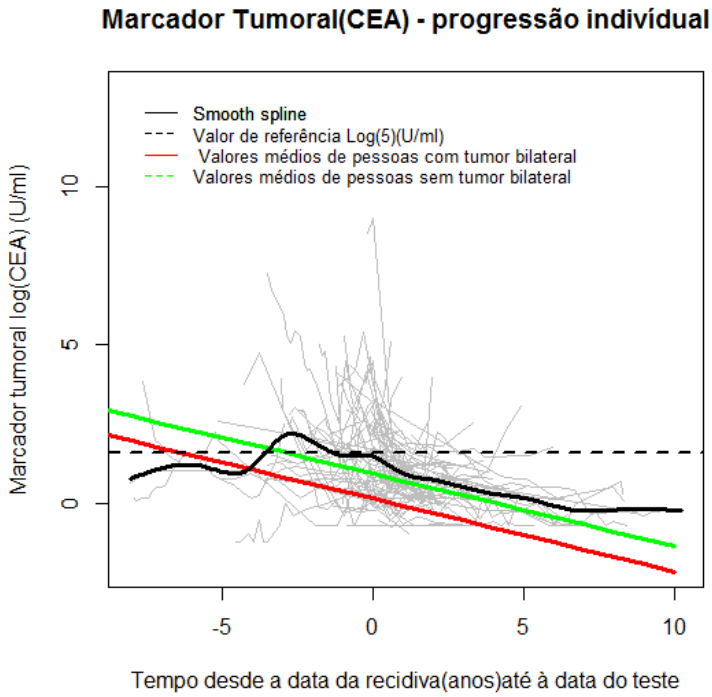



**Figura 58 - Sobreposição da progressão individual com a comparação das médias das pacientes com tumor bilateral**

Pela figura 58 podemos observar que uma paciente com tumor bilateral apresenta uma média do valor do marcador, mais baixo quando comparada com uma paciente sem tumor bilateral.

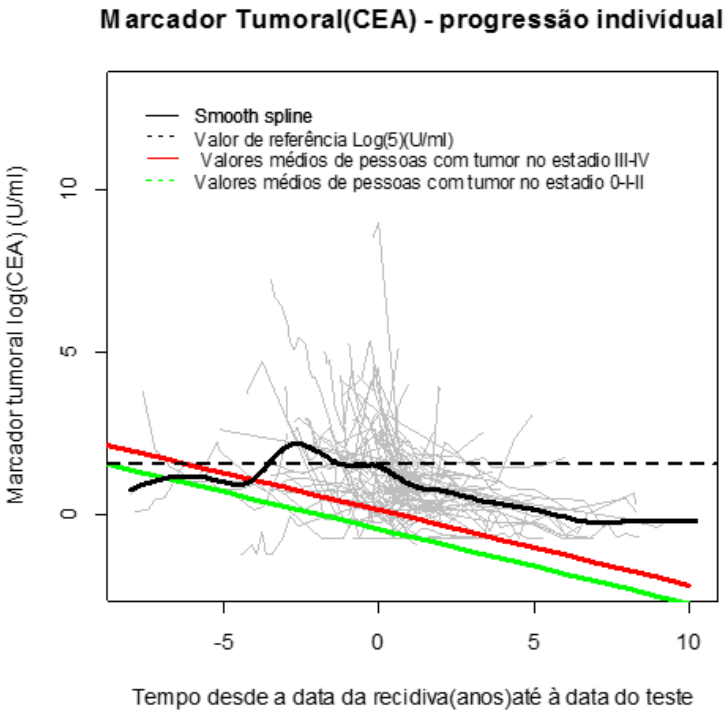



**Figura 59 - Sobreposição da progressão individual com a comparação das médias das pacientes nos diferentes estadios**

Na figura 59 podemos observar que uma paciente com um tumor no estadio III ou IV apresenta uma média do valor do marcador, mais alta quando comparada com uma paciente que com um tumor no estadio 0-I-II.

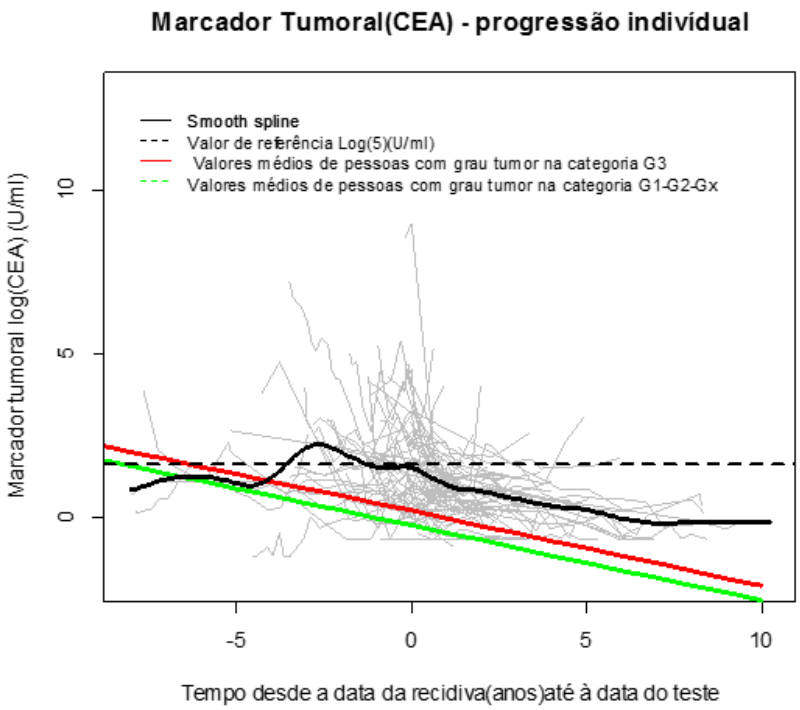



**Figura 60 - Sobreposição da progressão individual com a comparação das médias das pacientes com o tumor nas diferentes categorias**

Na figura 60 podemos observar que uma paciente com tumor na categoria G3 apresenta uma média do valor do marcador, mais alto quando comparada com uma paciente que tem um tumor na categoria $G_1$, $G_2$, Gx.

# Capítulo 7 - Conclusões

O interesse por esta temática deve-se à complexidade e mistério que ainda envolve uma doença que nos afeta, principalmente a nós mulheres e que, infelizmente, não obstante a ciência ter evoluído imenso na área da investigação, continua ainda a intrigar-nos pela evolução e desfecho que a própria doença envolve.

De forma a concretizar os objetivos do estudo aplicou-se inicialmente o modelo de Kaplan-Meier, onde o evento de interesse é a recidiva. O tempo de estudo é medido desde a data do diagnóstico até à recidiva.

Após a análise realizada, verificou-se que a sobrevivência para a recidiva a 120 meses é muito próxima de 80% e que para todo o seguimento observado a probabilidade de sobrevivência é superior a 40%.

Na análise de sobrevivência, foi ajustado um modelo de Cox incorporando múltiplos fatores de risco de modo a verificar qual ou quais tinham influência na probabilidade de ter recidiva.

Assim, as variáveis triplo negativo, tamanho do tumor, gânglios regionais e idade ao diagnóstico têm influência significativa no tempo até à recidiva, ou seja, uma paciente que apresente um tumor com triplo negativo, tem um risco de recidiva de cancro da mama aproximadamente 8 vezes superior a uma paciente que apresente um tumor sem triplo negativo; uma paciente com tumor na categoria $T_3$, $T_4$ ou $T_x$, tem um risco de recidiva de cancro da mama 4 vezes maior do que uma paciente que tenha um tumor na categoria $T_1$ $T_2$ ou $T_{is}$; o risco de recidiva de cancro em que os seus gânglios se encontram na categoria $N_3$, é 3,5 vezes maior do que $N_x$, $N_0$, $N_1$ou $N_2$; relativamente à idade ao diagnóstico o risco de recidiva é menor 60% para uma paciente com mais do que 44 anos, comparativamente a uma paciente cujo tumor foi diagnosticado com 44 anos, ou menos.

De modo a conhecer quais os fatores de risco que afetam a progressão ao longo do tempo de dois marcadores tumorais CA15.3 e CEA foram ajustados modelos longitudinais com diferentes estruturas de correlação.

É de salientar que estes dois marcadores têm valores de referência, e que se estes estiverem acima do valor correspondente é um alerta para uma possível recidiva do tumor.

Para o marcador CA15.3, em que o tempo de referência é a data desde o diagnóstico até à data do teste, as variáveis idade ao diagnóstico, imagens de invasão vascular venosa e Ki.67 são significativas na evolução do marcador.

A idade ao diagnóstico e imagens de invasão vascular venosa, provocam um aumento no valor inicial da progressão média do marcador tumoral contrariamente à variável Ki.67 que provoca uma diminuição no valor inicial da progressão média do marcador.

Podemos observar que uma paciente com 56 anos de idade diagnosticada com invasão vascular venosa e um Ki.67 baixo, apresenta uma média do valor do marcador mais alto quando comparada com um Ki.67 alto; e, uma paciente com 56 anos de idade diagnosticada com invasão vascular venosa e um Ki.67 alto apresenta uma média do valor do marcador mais alto quando comparada com uma paciente com Ki.67 alto mas sem invasão vascular venosa.

Relativamente ao tempo desde a recidiva até à data do teste, apenas a presença ou não de imagens de invasão vascular venosa tem influência na progressão do marcador. Uma paciente que tenha invasão vascular venosa tem um aumento nos valores iniciais do marcador tumoral de 3.9, quando comparado com uma paciente que não apresente vascular venosa.

Quanto ao marcador CEA, com o tempo de referência desde a recidiva até à data do teste, as variáveis bilaterais, estadio, grau e triplo negativo influenciam a sua progressão média significativamente.

No caso das variáveis bilateral e triplo negativo, provocam uma diminuição no valor inicial da progressão média do marcador CEA

No que concerne o tempo de referência desde o diagnóstico até à data do teste, as variáveis idade ao diagnóstico e estadio são significativas na progressão média do marcador CEA. Sendo que, estas duas variáveis provocam um aumento no valor inicial da progressão média do marcador tumoral CEA.

# Bibliografia

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. 50, 65

[2] Breslow, N. E. Covariance analysis of censored survival data. Biometrics, 30 (1):89-99, 1974.

[3] Bloom H, Richardson W "Histological grading and prognosis in breast cancer; a study of 1409 cases of which 359 have been followed for 15 years.". Br J Cancer 11(1957) (3): 359–77.

[4] Borges, A & Sousa I. (2015). Joint Modelling of Longitudinal and Survival Data on Brest Cancer. Universidade do Minho, Braga, Portugal.

[5] Cianfrocca, M. & Goldstein, L. (9). Prognostic and predictive factors in early stage breast cancer. The Oncologist, 6. 15,17, 31

[6] Cox DR. Regression models and life-tables, Journal of the Royal Statistical Society, series B, 34, 87-220, 1972.

[7] Cox, D. R., & Oakes, D. (1984). Analysis of Survival Data. London, N Chapman and Hall, London – New York 1984.

[8] Cox, D. (1975). Partial likelihood.Biometrika, 62, 269–276. 47

[9] Cressie, N. (1993). Statistics for Spatial Data. Wiley, New York.

[10] Diggle PJ; Heagerty P; Liang K-Y, and Zeger SL, (2002), Analysis of Longitudinal Data; University Oxford Press

[11] Diggle, P., Heagerty, P., Liang, K. & Zeger, S. (2002). Analysis of Longitudinal Data. Wiley, 2nd edn. 73, 74, 77, 78,79, 80, 81, 118, 142

[12] Ferreira, J. M. (2007). Análise de Sobrevivência: uma Visão de Risco Comportamental na Utilização de Cartão de Crédito. Dissertação (Mestrado em Biometria e Estatística Aplicada) - Universidade Federal Rural de Pernambuco.

[13] Fitzmaurice, G., Davidian, M., Molenberghs, G. & Verbeke, G. (2008). Longitudinal Data Analysis. CRC Press. 74,75, 76models. Biometrika, 73, 13–22. 75, 76

[14] Hair, J.F. Jr., Anderson, R.E., Tatham, R.L., & Black, W.C. (1998). Multivariate Data Analysis, (5th Edition). Upper Saddle River, NJ: Prentice Hall.

[15] Kaplan, E. & Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53, 457–481. 43

[16] Liang, K.& Zeger, S. (1986). Longitudinal data analysis using generalized linear

[17] Mason R. L., Gunst, R. F.& Hess, J. L. (2003). Statistical Design and Analysis of Experiments: With Applications to Engineering and Science. J. Wiley, New York

[18] RORENO 2000/2001, Análise de Sobrevivência – Principais Cancros da Região Norte 2000/2001.

[19] Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. Biometrika, 69, 239-41.

[20]http://justnews.pt/noticias/unidade-de-senologia-do-hospital-de-braga-luta-contra-falta-de-referenciacao#.WIX3YNKLTIU, acedido em 10 de dezembro de 2016.

[21] http://globocan.iarc.fr/Default.aspx, acedido em 12 de dezembro de 2016.

[22] http://gco.iarc.fr/today/home, acedido no dia 19/07/16